Summary of Public Feedback on the 2010 Demonstration Data Product - Demographic and Housing Characteristics File

(2022-08-25)

Written By: Jason Devine and Alexandra Krause

Summary

For the 2020 Census, the Census Bureau is using a new framework for disclosure avoidance for the decennial data products. This new methodology works by adding small amounts of noise – or variation - to the published data. By applying the latest iteration of the disclosure avoidance system to 2010 Census data tables, the demonstration data products allow data users to observe and assess the amount of noise that would be added to the 2020 Census data products and to compare the 2010 noise-infused data to the published 2010 Summary File 1 data, which were protected using swapping. The Census Bureau released two demonstration data products that supported all the proposed content for the 2020 Demographic and Housing Characteristics File (DHC). For each demonstration data product, the public had a minimum of 30 days to review and provide comments. This document summarizes the feedback on the second 2010 Demonstration Data Product - DHC released on August 25, 2022.

The Census Bureau received 11 comments during the public comment period for the second round of **2010 Demonstration Data for the Demographic and Housing Characteristics File (DHC) (v. 2022-08- 25)**. Of those, 10 included feedback related to the accuracy of the data. All comments were provided through the 2020DAS@census.gov email address. Additional feedback was provided through engagements with stakeholders and Census advisory groups, which is not contained in this summary. The written comments are summarized here, and the full text of comments are included in Appendix B.

Many commenters acknowledged improvements in the second round of demonstration data, however, there were still numerous concerns. The accuracy of age data was still identified as the top concern even for geographies as large as counties. Commenters continued to be concerned about the accuracy of tenure and other household data (e.g., household type, household size, multigenerational by race and ethnicity, and vacancy). They were particularly concerned about the relative error for small population groups, such as age or tenure by race and ethnicity. Despite improvements since the first demonstration product, inconsistencies and improbable results remained a concern.

Background

The U.S. Census Bureau has a long history of using disclosure avoidance techniques to protect the confidentiality of respondents' personal information, which is mandated by law.[1] Over time, advances in computing power and availability of commercially available databases on people and households increased the risk that individuals could be reidentified from published statistics. To mitigate this risk,

[1] U.S. Constitution, Article I, Section 2; Title 13 U.S. Code, Sections 8–9; Title 13 U.S. Code, Section 141.

the Census Bureau adopted differential privacy as its framework of disclosure avoidance for the 2020 Census, applying a formal privacy mechanism for the first time.[2] As with any disclosure avoidance system, the use of differential privacy involves tradeoffs between the accuracy and granularity of the published counts and the extent to which confidentiality is protected. As a result, the Census Bureau has conducted extensive outreach with the public to determine a balance of accuracy and privacy, including the release of demonstration data products.[3]

Demonstration data products allow the public to assess the impact of differential privacy on accuracy by applying the latest iteration of the disclosure avoidance system to the 2010 Census data. The public can compare the demonstration data, protected with differential privacy methods, to the published 2010 Census data, protected with swapping. This comparison is not perfect. Differential privacy is applied to the pre-swapped 2010 Census data, so some differences with the published data could be due to the 2010 Census disclosure protections that relied on swapping. Demonstration data products are in summary file format and accompanied by Detailed Summary Metrics – an Excel-based document that shows measures of differences for a variety of topics and geographies. The Census Bureau sought public feedback on whether the level of noise was acceptable. If it was not, the Census Bureau asked for the impacted table or topic, geography, and use case. Because the 2020 DHC is expected to be consistent with the 2020 Redistricting data (Public Law 94-171), the DHC demonstration data were held consistent with the production-settings demonstration data for the Redistricting Data Summary File. This includes occupancy status, total population, Hispanic origin, race, people under and over 18 years, and major group quarters type.

During early stages of DHC disclosure avoidance development, the Census Bureau released two demonstration data products that contained limited DHC content. These data products were released in October 2019 and May 2020. The Census Bureau then paused work on the full DHC to focus resources on the pending Redistricting data release. They resumed DHC development following that release in August 2021 and produced two demonstration data products that contained all proposed DHC content. We refer to these as the first and second round demonstration data products.

The first round of demonstration data released the person and housing unit files separately. The person file includes person-level characteristics (e.g., Hispanic origin and race). The housing unit file includes characteristics of the housing unit and householder (e.g., tenure and race of householder). The set of person-based tables were released on March 16, 2022. The housing unit-based tables were released on March 29, 2022 and re-released on April 14 to correct a technical error. The public comment period was March 16 to May 16, 2022, and the Census Bureau received 14 comments. These comments were previously summarized.[4] The second round of demonstration data were released as one set of person

[2] For more information on differential privacy, see U.S. Census Bureau, Disclosure Avoidance for the 2020 Census: An Introduction, U.S. Government Publishing Office, Washington, DC, November 2021.

[3] Note, the Census Bureau collected feedback on the DHC data product proposal in addition to the DHC disclosure avoidance system. Notably, there was a 2018 Federal Register Notice (FRN) for 2020 Census data products and a call for feedback on the 2020 Census Data Product Planning Crosswalk in October and December 2021. These calls for feedback were focused on the design and content of the DHC rather than accuracy of the data. Feedback helped inform the proposed design of the DHC.

[4] Summary of Public Feedback Received on the 2010 Demonstration Data Product – Demographic

and housing unit tables on August 25, 2022. Public comments were due on September 26, 2022, and the Census Bureau received 11 comments. This paper summarizes these 11 comments.

Development of the disclosure avoidance system to support the DHC has been an ongoing process; the Census Bureau subject matter and differential privacy experts develop and analyze experimental files throughout the feedback periods. After the second demonstration data product was released, the Census Bureau continued to make improvements during and after the public comment period. The Census Bureau will demonstrate these improvements to the public via a production settings demonstration data product for the DHC, similar to what was produced for the Redistricting data.

Comments by Type of Affiliation

The second round of demonstration data product comments represented a cross-section of data users, including federal, state, and local governments, academic institutions, private companies, and nonprofit organizations (see Table 1). State governments provided the most comments (4 comments) followed by local governments and non-profit organizations (2 comments each). Federal government, academic institutions, and private companies submitted one comment each.

Table 1. Comments by Type of Affiliation

Type of Affiliation	**Count**
Federal Government	1
State Government	4
Regional Government	0
Local Government	2
Academic Institution	1
Private Company	1
Non-Profit Organization	2

Geography

Commenters referenced a range of geographic levels (see Table 2). Although tract data was mentioned the most, commenters continued to be concerned about the accuracy of census block data while acknowledging the Census Bureau's guidance of aggregating single blocks, the lowest census level of geography, with other blocks. They warned that many data users would continue to use block-level data regardless of the guidance. One commenter concluded that even aggregated block data were not useable in many cases.

Commenters mentioned tract the most (7 comments) followed by census block (6 comments) and county (4 comments). In a slight shift from the first round of demonstration data feedback, there was more concern about tract and county level accuracy in the second round of demonstration data feedback. Commenters were also concerned about the accuracy of off-spine geographies, including incorporated places (2 comments), unincorporated places (1 comment) and school districts (2

and Housing Characteristics File (2022-03-16) at https://www2.census.gov/programs-surveys/decennial/2020/program-management/round_1_feedback.pdf comments).[5] Many commenters noted that accuracy for small populations was a concern even at larger geographic levels, such as counties.

Table 2. Comments by Geography

Geography	Count
Census Block	6
Tract	7
County	4
Incorporated Places	2
Unincorporated Places	1
School Districts	2
Off-spine Geographies	2
State	1
Neighborhood Tabulation Areas (NTAs)	1

Commenters' Concerns

While fewer comments were received on the second-round demonstration product (11 comments), compared to the first-round demonstration product (14 comments), the number of commenters' concerns about accuracy was similar – 10 and 11 respectively. More specifically, nine commenters were concerned about accuracy for specific topics or geographies, five were concerned about accuracy for small populations or small geographies, three were concerned about bias, and three were concerned about inconsistent or improbable results. These concerns were not mutually exclusive as many commenters had multiple concerns. Two commenters were concerned about 2020 Census data product proposals. In addition, one commenter was concerned about the decline of public trust in the Census Bureau, and another cited a concern about equity of data accuracy.

One commenter was concerned about privacy. In response to the request for feedback on the first-round demonstration data product, they analyzed the potential disclosure risk of identifying transgender children. They re-did the same study for the second round of demonstration data and reached a similar conclusion. The study found that differential privacy provides more confidentiality protection than previous disclosure avoidance methods, such as swapping. See Table 3 for an overview of commenter concerns.

Table 3. Commenter Concerns

Concern	Sub Concern	Count
Accuracy		10
	Accuracy for Specific Topics or Geographies	9
	Accuracy for Small Populations	5
	Bias	3
	Inconsistent/Improbable Scenarios	3

[5] Off-spine geographies are those geographies that do not nest within the Census geographic hierarchy that includes the nation, regions, divisions, states, counties, tracts, block groups, and blocks.

Data Product Proposal		2
Public Trust		1
Equity		1
Privacy		1

As previously mentioned, ten commenters were concerned about accuracy overall. Most were concerned about several topics, and multiple commenters mentioned the same topic (see Table 4). Eight of the ten accuracy comments cited concerns about sex by age data, and most of these were also concerned about sex by age by race and ethnicity. In particular, five comments cited concerns about the accuracy of sex by selected age categories, two cited concerns about median age, two cited concerns about single-year of age, and two cited concerns about children (under 18 years and 0-4 years). Concerns about the sex by age data were focused on large relative or percent differences for smaller populations—even for larger geographic areas, such as counties.

The second most prevalent concern was about tenure crossed by other variables. Four comments cited concerns about the accuracy of tenure data, specifically by race and ethnicity, presence of children, household type, and household size. The following topics received two comments each: race of householder, household type, and household size. Multigenerational household by race and ethnicity and vacancy received one comment each. Commenters tended to rely on aggregate measures and distributions of numeric and percent differences rather than differences for individual geographic entities. Some of these concerns may have been alleviated or resolved as the Census Bureau continued to make improvements to the disclosure avoidance system after releasing the second round of demonstration data.

In addition, some commenters cited concerns about data that had already been released in the Redistricting data product and will remain unchanged in the DHC (i.e., total population, race and ethnicity, household population, group quarters population, and occupancy status).

Table 4. Concerns by Topics and Subtopics

Topic	**Subtopic**	**Count**
Sex by Age		8
	Sex by selected age categories	5
	Median age	2
	Single-year of age	2
	Children under 18	1
	Ages 0-4	1
Tenure crossed by other variables	Race/ethnicity, presence of children, household type, and household size	4
Total population		3
Race and ethnicity		3
Impossible/improbable scenarios		3
Household population		2
Group quarters population		2
Race of householder		2
Household type		2
Household size		2

Multigenerational household by race and ethnicity		1
Occupancy status (occupied and vacant)		1
Vacancy		1

In addition to specific topics, commenters were still concerned with inconsistent and improbable results. An inconsistent result can occur when combining tables produced from the person and housing unit files. This inconsistency represents the uncertainty that the disclosure avoidance deliberately introduced when it was applied independently to the person and housing unit files with no mechanism to merge the files or added constraints to maintain consistency. This led to inconsistent and improbable results when comparing across tables. For example, race on the person file may not be consistent with race of householder on the housing unit file. Impossible and improbable results occur more often for smaller levels of geography, such as blocks, and less frequently for larger geographies. Commenters warn that these inconsistent and improbable results will undermine the data users' confidence in the data and may lead them to conclude that the data are unusable.

Concern about the Accuracy of Sex and Age Data

Commenters focused heavily on the sex by age data. They explained that age data are used in applications with many downstream impacts. For example, age data are used for population estimates, which are used as controls for various surveys and denominators in rates. With population estimates traditionally relying on census data as their base, inaccuracies will be carried through the decade as each age cohort is aged forward each year.

In addition to data on children, commenters mentioned 5-year age groups, age-specific and age-adjusted rates, and impossible and improbable scenarios related to children. After recreating an earlier analysis with the new demonstration data, a commenter reported some improvements but was still concerned about the usability of county and tract data for age-specific and age-adjusted rates. They examined the impact on rates developed for several programs at the county and tract levels. For each of the programs, they concluded there would be a "profound" and "significant" impact to the results.

Many commenters referenced geography in relation to age data, including off-spine geographies, less populous geographies, and larger geographies. Based on a commenter's analysis, accuracy was worse for off-spine geographies, compared to on-spine geographies. In addition, accuracy was worse for geographies further from the spine (e.g., unincorporated places), compared to those closer to the spine (e.g., incorporated places).

Comments cited concerns for larger geographies that were less densely populated where data users "expected better" accuracy. These findings could represent disparities in the usability of age data where populous groups and geographic areas have more usable data for lower geographic levels, such as tracts, and less-populated groups and geographic areas have less usable data for larger geographies, such as counties, particularly when examining percent differences. For one commenter, this led to a concern about equity.

Commenters were concerned about age data for even larger geographies, such as counties. Even at the county level, groups can be small when parsing sex by 5-year age groups by race and Hispanic origin. With these small populations, small numeric differences can result in large percent differences. Commenters warned that these larger relative differences for small populations introduced uncertainty into the data and would impact its use for a variety of purposes.

Concerns about Tenure and Other Housing Unit Data

Tenure by other variables (e.g., race and ethnicity, presence of children, household type, and household size) was the second largest concern. One commenter noted bias for tenure by age of householder. For example, the smallest categories (i.e., owner-occupied households among young adults) tended to gain population. One commenter reported that the share of rented to total housing units was different for nearly all block groups and tracts in Virginia. When analyzing the same table for a subgroup of the population, they found a substantial number of unusable results at the tract level. For example, Hispanic homeownership experienced a large change where 72 out of 520 tracts (14 percent) had differences equal to or greater than 10 percent. For one tract, the Hispanic homeownership value grew from 21 in the published 2010 Census data to 38 in the demonstration data. This commenter considered differences greater than 10 percent unusable and differences less than 5 percent acceptable. They stated, “this level of error makes this data very inaccurate and unusable for government policy-making and service planning efforts.”

Commenters cited other housing unit data, including household type, household size, multigenerational households by race and ethnicity, and vacancy. One commenter noted lower accuracy for less common household types and smaller households. They explained, “table cells for household type and household size that generally have lower counts (less common household types and household sizes) often have very large percentage errors (e.g., over 30 percent of tracts have more than 10 percent error), which severely limits the usability of these tables.”

Commenters mentioned multigenerational households in the content of diversity indices. Mainly, demonstration data tend to increase racial and ethnic diversity. They explain, “diversity indices in the demonstration data are higher than in [the published 2010 Census data], especially for the rarer category of households with three or more generations.” One commenter provided an analysis for a variety of housing unit variables that they frequently use for research and policy decision making, including vacant units for seasonal, recreational, or occasional use. For one block group, the vacant unit count for seasonal, recreational, or occasional use was 28 in the published 2010 Census data and decreased to 4 in the demonstration data—an absolute difference of 24 vacant housing units or 600 percent. The commenter concluded, “there were a large number of [topics] where the modification was very significant.”

Concerns about Impossible and Improbable Scenarios

Some data inconsistencies and improbable scenarios exist because disclosure protections are applied to the person and unit files independently. These inconsistencies are a feature of the disclosure avoidance system. While commenters generally understood that these inconsistent and improbable scenarios could not be eliminated, almost half cited concerns about the number of them. They explained, “of further concern are the number of blocks that have fatal inconsistencies which would invalidate any attempt to use the data to build either a custom geography (say a special district) or in tracking disasters, or public health issues.”

During the first round of demonstration data, commenters noted concerns about these inconsistent and improbable scenarios, and they continued to cite concerns in the second round of demonstration data. Examples of these scenarios include:

- More households than household population
- Household population without households
- Householders not equal to households
- Household population under 18 less than the number of households with children under 18
- Not enough population to match a population calculated from household size
- More householders of a certain age than population of that age
- Blocks with only children under the age of 18 years

Commenters continued to find that “block-level data is full of inconsistencies” with some inconsistencies occurring less frequently as the population size of an area increased. Other inconsistencies were pervasive, impacting almost every geographic unit. An example of a pervasive inconsistency was the number of householders compared to the number of households.

Concerns about Bias

One commenter identified many biases in the first demonstration product. The same commenter noted something similar for the second round of demonstration data. They explained that various biases exist within the demonstration product that are not well-understood. For example, “areas with mostly renter-occupied houses have a positive bias for households with children, whereas areas with mostly owner-occupied houses have a negative bias for households with children.” They explained it will take additional analysis to better understand these biases, and by that time, the data will already be released.

Acceptable Accuracy

Data that were considered fit-for-use in the first demonstration product were still considered fit-for-use in the second demonstration product. In addition, many commenters acknowledged improvements from the first round of demonstration data. For example, many acknowledged improvements in age data. Some noted 5-year age data were fit-for-use for total population and household population. Another noted age ranges 0-4 as fit-for-use for states and large counties. One commenter noted sex by age for total population for on-spine and near-spine geographies with 1,000+ people was fit-for-use. Finally, one commenter mentioned that spatial clustering for total population was fit-for-use. These findings emphasize that data were generally acceptable if the group or geography was populous.

Some commenters provided thresholds for fitness-for-use. One commenter considered a percent and numeric error of 5 or less as acceptable. Another considered a difference of three percent for a 5-year age group problematic. One commenter thought that differences should not exceed 100 percent even if the differences only occurred for small population groups. Taken together, the comments start to provide a rough gauge for what could be considered acceptable.

The Census Bureau notes that average percent differences need to be considered with caution. Cells with small numeric differences can skew the average percent difference creating the perception that many cells exhibit a high percent difference. While commenters acknowledged this when providing their comments, they did not always take steps to account for the impact of zero and small cells in the measures of differences their conclusions were based on.

Recommendations

Table 5 summarizes commenter recommendations. Most comments requested additional accuracy for specific topics and/or geographies (9 comments), particularly additional accuracy for small populations and/or small geographies (5 comments). In addition, many commenters recommended reducing bias (3 comments) and impossible and improbable scenarios (2 comments). Two commenters requested updates to the 2020 Census data product proposals. Finally, the following recommendations received one comment each: the Census Bureau should develop privacy metrics, so the public can truly assess the accuracy and confidentiality trade-off; the Census Bureau should work to produce equitable data; and the Census Bureau should use differential privacy, compared to previous disclosure avoidance methods, such as swapping. Some of the same recommendations were included in the first round of demonstration data, and the Census Bureau has worked to make improvements in these areas.

Table 5. Commenter Recommendations

Recommendation	Count
Increase accuracy for specific topics and/or geographies	9
Increase accuracy for small populations and/or geographies	5
Reduce bias	3
Reduce impossible/improbable scenarios	2
Update data product proposal	2
Develop privacy metrics	1
Produce equitable data	1
Use differential privacy	1

Compared to the first round of demonstration data, commenters described the potential impact if the data were released as-is. They cited negative impacts on population estimates and projections, funding for programs, city and neighborhood planning (e.g., school-related needs), informed decision making (e.g., policy decision making), and health disparities (e.g., prevalence and progression of cause-specific mortality). In addition, commenters noted the potential impact of the Census Bureau losing public trust. For specific quotations on impacts, see Appendix A.

Conclusion

Overall, the commenters continued to recommend increased accuracy for age, tenure, and other household data (e.g., household type, household size, multigenerational by race and ethnicity, and vacancy)—both alone and when crossed by race and ethnicity. The one commenter that considered privacy by examining the potential disclosure risk of identifying transgender children concluded the privacy settings were effective. Commenters continued to cite concerns about the frequency of inconsistent and improbable results. These comments suggest that as of the second demonstration product, improvements were still needed for age and housing unit data more generally. Greater accuracy across the housing unit file not only improves the accuracy of the housing tables, but it also reduces the number of person and housing unit inconsistencies. The Census Bureau worked throughout the demonstration product comment periods to increase accuracy within privacy constraints. The Census

Bureau plans to release a production settings demonstration data product for DHC, which will allow data users to assess the final result of the disclosure avoidance protections. We thank those who provided feedback, which has been an instrumental part of the developmental process.

Appendix A: Selected Quotations Regarding Impacts

Errors of the magnitude shown above could have important implications for federal and state funding received by schools and for educational planning. Errors of this magnitude might impact formula funding that is based on Census-derived data and some schools would get less than they deserve.

First, blocks with children and no adults are a highly implausible situation and the large number of such blocks may undermine confidence in the overall Census results.

Reamer (2020) shows that $39 billion of federal funds were distributed by the U.S. Department of Education to states and localities in FY 2017 based on census-derived data. Table 2 shows programs run by the U.S. Department of Education that distribute federal funds to state and localities based on census-derived data... Overall, Reamer (2020) identified 316 federal programs that use census-derived data to distribute about $1.5 trillion to states and localities in Fiscal Year 2017. About two-thirds of the 315 programs use substate data which underscores the importance of small area census data. When one is talking about billions of dollars, a small percent error can translate into a large dollar amount.

Current and projected demographic data are often used to construct attendance boundaries to keep classrooms from becoming overcrowded. Constructing attendance boundaries often include sensitivity to racial composition, so small area demographics by race are important. Such activities often require very small area data such as census blocks.

Cities, villages, and towns might want to know about the number of young children in their area for things like planning youth activities, child facilities, and day care centers. The preschool-age population is also useful for forecasting future school enrollments.

Consequently, census accuracy for blocks is especially important. O'Hara (2022) makes a strong case for why block level data are important in terms of creating special or custom districts. The need for such data is often not apparent until well after the Census data has been collected and reported.

Spatial clustering of population underestimates would result in underfunding or underrepresentation for programs based on population counts. The impact of large over or underestimates is especially problematic where population totals are small. Some of the absolute errors are as high as 2245 percent. This seems unacceptable.

Having accurate county level data for projection is crucial as most MCDs in Minnesota, as I have pointed out to the Bureau on multiple occasions, have less than 1,000 residents. Given the issues presented here, I have no confidence in the ability of researchers and data users to be able to produce subcounty *projections for anything other than our few largest cities. This will be due to the intentional perturbation of the 2020 census data.*

When we look at the larger picture, the data provided above show that 44 counties in the state will have insufficient data to produce population projections with a cohort component method. That represents over 50 percent of the counties in the state. That cannot be acceptable to the Bureau. Governmental units need these data for planning and resource allocations regardless of the size of the area in question.

The errors are so large that they could have important implications for federal and state funding received by schools and for educational planning. Errors of this magnitude might impact formula funding that is based on Census-derived data such that some schools would get less than they should by law. It could also distort demographic predictions of school population, affecting plans for school buildings and class size. We urge the Census Bureau to try and reduce or eliminate these large errors.

In New York City, it is not enough to know, for example, that the Asian population has decreased in Manhattan's Chinatown. We must disentangle subgroup information by race, distinguishing whether it was the Chinese or Vietnamese population that declined in this example, so that we can properly allocate resources for services that our residents require... The Census Bureau has invested years of work towards improving and expanding the race and Hispanic questions, so that we can more precisely portray our increasingly diverse population. ... However, the current product plan does not do justice to this collection effort, and respondent burden, because it fails to tap the rich responses that can better portray the diversity of neighborhoods across the nation.

I am concerned that the level of error contained in the current DHC demonstration file may negatively impact our Department's ability to measure and address health inequities if it is retained in the final DHC file, as it will not allow us to access highly accurate population estimates by broad race and ethnicity categories at sub-county geographies for LA County.

Hawaiian/Pacific Islander (NHPI) populations (two priority populations for our Department). To provide a recent example, we have relied on having access to accurate census-tract level race and ethnicity population data throughout the COVID-19 pandemic to understand the disproportionate impact of the pandemic on our county's AIAN and NHPI communities. These data have also been used extensively to inform our COVID-19 vaccination efforts, from estimating coverage rates in various subpopulations by city and community to informing our targeted outreach efforts

---

DHC has historically served as the foundation for the detailed postcensal population estimates that are procured by the Los Angeles County government for several of its departments, including the Department of Public Health. These postcensal estimates necessarily provide highly detailed demographic information at the split census tract-level (stratified by detailed age groups, race and ethnicity, and sex). Our Department relies on these postcensal estimates as a source of information during intercensal years for the demographic composition of the county overall and various sub-county geographies as well as a *source of denominator data to use for our incidence and prevalence rate calculations. Errors in the DHC file will therefore be carried forward for the remainder of the decade when these subsequent postcensal estimates are produced. . . We are concerned that if the level of error introduced by the current tuning of the DAS in the DHC demonstration data remains in the final DHC file, the subsequent population data that we frequently use to characterize or county populations at various sub-county geographies or as denominators in our incidence and prevalence rate estimates could be distorted.*

In light of recent laws prohibiting parents from obtaining medical care for their trans children, our results demonstrate the importance of disclosure avoidance for census data, and suggest that the TopDown approach planned by Census Bureau is a substantial improvement compared to the previous approach, but still risks disclosing sensitive information.

Appendix B
Full Text Compilation of Feedback Received

2010 Demonstration Data Demographic and Housing Characteristics File (DHC) v. 2022-08-25

Round 2 Feedback

2010 Demonstration DHC #2 Feedback – Internal

1. Thomas F Petkunas, Census Bureau Employee

We have attached feedback for the demonstration data released on 8-25-2022 for the 2020 Census Demographic and Housing Characteristics File (DHC).

Thank you for your attention.
-Tom

Comments for DHC Demonstration Data

9-26-2022

General Remarks:

Overall, working with the DHC demonstration data is a bit challenging and possibly beyond the ability of the general public users. Prerequisite knowledge of other Census data sets, such as SF1, is needed to access and analysis the data. Our comments and recommendations for the DHC demonstration data deal more with accessing and using the data sets to produce summary statistics, rather than the actual data/numbers themselves.

Incorporated improvements:

Improvements/updates have been made in the past that allowed the user to more easily access and work with the data. One of those improvements to the DHC data sets that we found extremely helpful was that the data sets were now field delimited. This drastically cut down the time it took to incorporate data into any desired software (R, Excel, Oracle, etc) and, more importantly, eliminated errors.

Geography files:

We wanted to make comparisons of the DHC data with the SF1 data. Our desire was to compare data at multiple levels, starting at the county level and building off that hierarchy. We were hoping to see a true GEOID field, whether it was 15-characters or 17-characters. A GEOID field that was consistent with other Census data products.

A dedicated column for the 15-digit GEOID, and renaming the current 60-char GEOID field, would be useful. One needs to parse together these elements (State, County, Tract, Block) to create this identifier. This is not a cumbersome task by any means, but it leads to some initial confusion.
The current GEOID & GEOCODE fields are confusing and are, respectively, 60 & 51 characters long. The user needs to guess as to which characters need to be parsed together to create the 15-char GEOID, because the documentation does not include a char-by-char description of these fields.
There are STATE, COUNTY, TRACT, BLOCK fields available, but they aren't always complete.

Targeted Audience:

Who is your target audience? First time users of the data and data sets can find it confusing and frustrating. We have been currently working with these data sets for 6 months. After reading the entire technical documentation and experimenting with the data sets, we feel as though we finally have a good grasp on the data structure, however, we are making a lot of assumptions that our understanding of the structure is correct.

For example, we tried to create and store our own 15-character GEOID by concatenating the STATE, COUNTY, TRACT, BLOCK fields. Of course, the number of counties we came up with via this method was much greater than the actual Census counts. It took us a while to determine that we also had to look at other fields (ie, LSADC) to narrow down our scope.

2. Dan Barroilhet, Demographer Division of Intergovernmental Relations, State of Wisconsin

Dear Census colleagues,

My job requires me to develop annual population estimates and periodic population projections for each Wisconsin county and county subdivision. I help businesses, and state/county/municipal officials find and interpret data from the Census Bureau and partner agencies (primarily BLS and BEA). I'm an active member of the State Data Center network and the Federal State Cooperative for Population Estimates and the Federal State Cooperative for Population Projections.

For my work, I would allocate 100% of the privacy budget to a few essential places. Occupancy/vacancy. Household size (household population per occupied housing unit). Female age distribution by 5-year age group (0-4, 5-9, etc.) Male age distribution by 5-year age group. My work requires no detail about race or ethnicity.

Except Regional Planning Commissions, almost no data user can use multi-county data. At least 90% of data users start with county or county subdivision data and need to drill down to tract or block to get their work done. For this reason, I consider the county to be the "largest" geographic unit that users refer to with any frequency. Wisconsin's 2010-2020 ten-year change was less than 4% and many (perhaps most) Wisconsin counties will have 2020-2030 change less than 2%. For this reason, I use 2% means absolute percent difference as a threshold of high concern.

The attached PDF shows mean absolute percent differences for six Wisconsin counties' sex-specific age distributions. The difference is the absolute value of (PP DD – SF1)/SF1. The first page shows Burnett County males' mean absolute percent difference of 2.61%. The second page shows Florence County (female MAPE = 2.91%; male MAPE = 4.86%). The third page shows Forest County (male MAPE = 3.45%). The fourth page shows Menominee County (female MAPE = 5.56%; male MAPE = 10.85%). Also, female percent of total pop MAPE = 4.71, with a rather different pattern by age groups). The fifth page shows Pepin County (female MAPE = 2.0%; male MAPE = 3.31%). The sixth page shows Sawyer County (female MAPE = 2.45%; male MAPE = 1.97%).

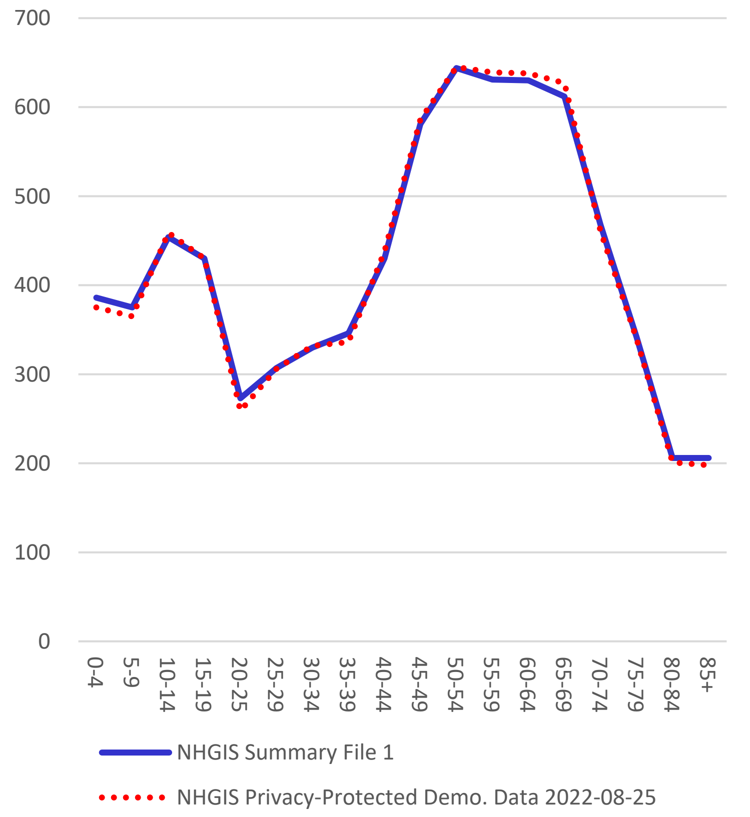
Say a county has to allocate public health spending between something like SIDS, which affects younger residents, and something else like Alzheimer's, which affects older residents. A 2% MAPE for age data is a really significant. For population projections, it's necessary to calculate age-specific fertility (for females) and age-specific mortality (for females and for males). If females' age distribution is off by an average of ≥ 2% in each age category, this really jumbles a few things. (1) The base is off and the erroneous distribution gets aged forward through the entire projection period. (2) Fertility calculations are distorted by age distribution errors. (2) Mortality calculations are distorted by age distribution errors. (3) Migration calculations are distorted by age distribution errors.

Yes, I do understand that 2010 Summary File 1 used swapping to protect privacy, so it is not error-free. Technically, it's more correct to say mean absolute percent *difference*. However, it's worth remembering a couple of things. (1) SF-1 preferred to swap within the tract, when possible, and almost always swapped within the county, so county figures wouldn't be affected. (2) Swapping will more often change a characteristic like race or ethnicity; age and sex will often be held constant.

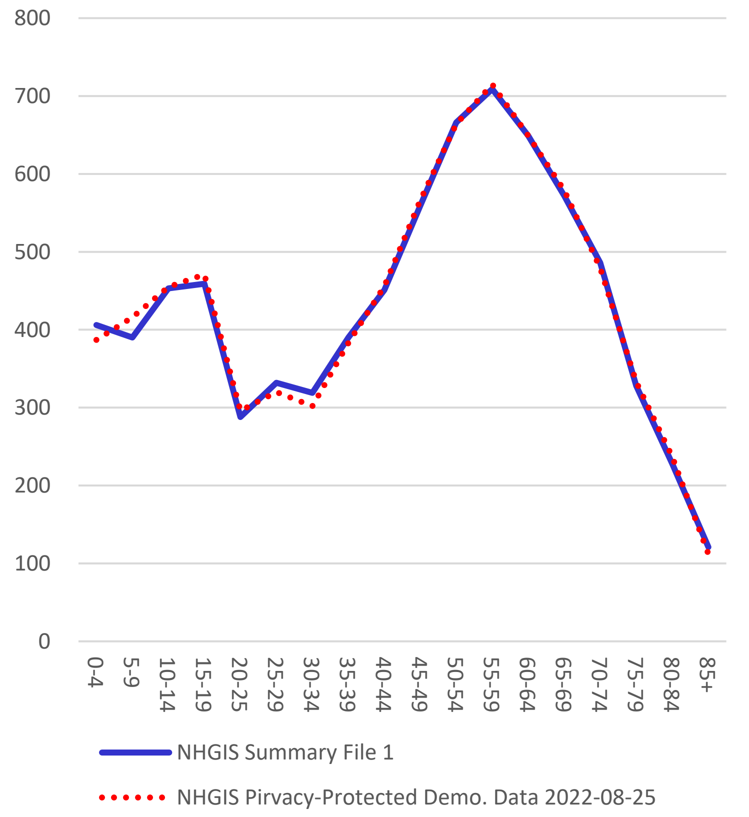
Yes, I do understand that the current iterations demonstration data reflect a lot less error than previous iterations. Say I have to retake calculus. The professor tells me that after I retake the class, my current rate of improvement would get me to a passing grade in 23 years. My second retake (my third time through the class) suggests I'll get a passing grade in 17 years. This is very significant improvement, but I'm still a long, long way from a passing grade. Two things can be true at the same time: DAS has improved very significantly; the rate of improvement still will not achieve a passing grade if 2020 Census operations conclude before 2032. Please keep up the good work.

Dan Barroilhet
Demographer, Research Analyst
Division of Intergovernmental Relations
State of Wisconsin

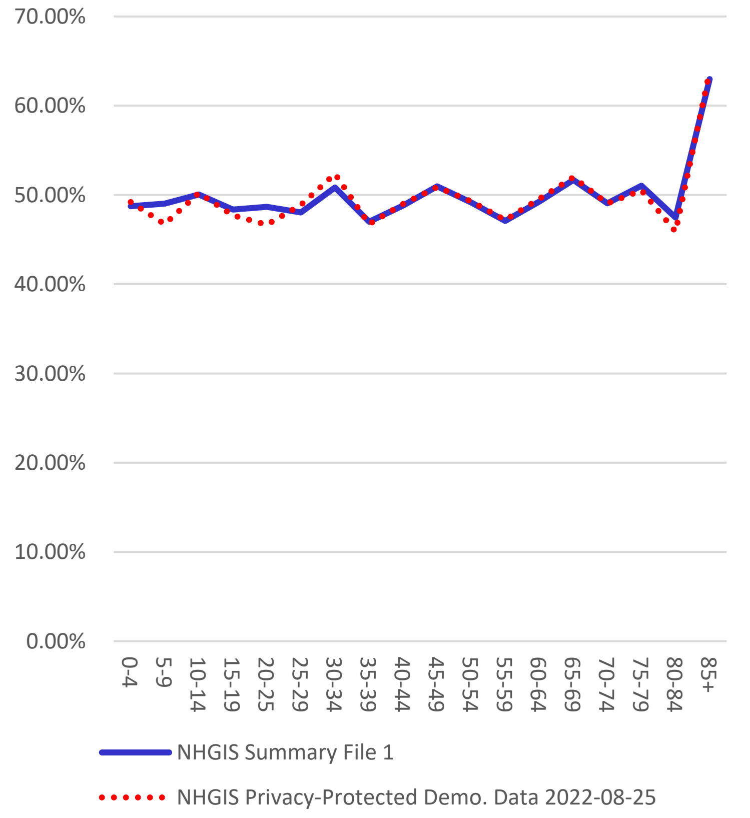
Female Age Distribution
Burnett County, Wisconsin
mean abs % diff = 1.77%



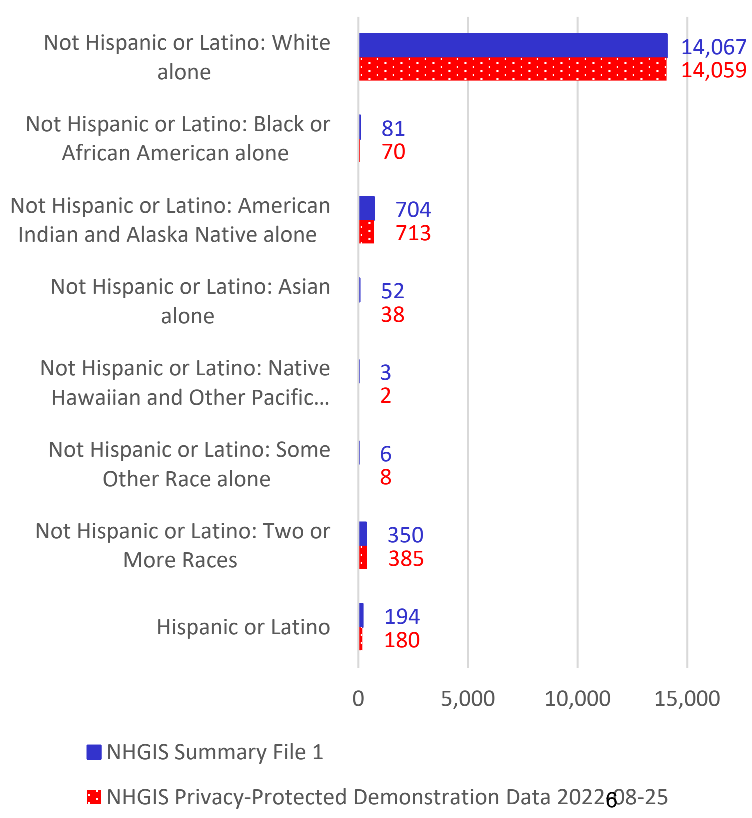
Male Age Distribution
Burnett County, Wisconsin
mean abs % diff = 2.61%

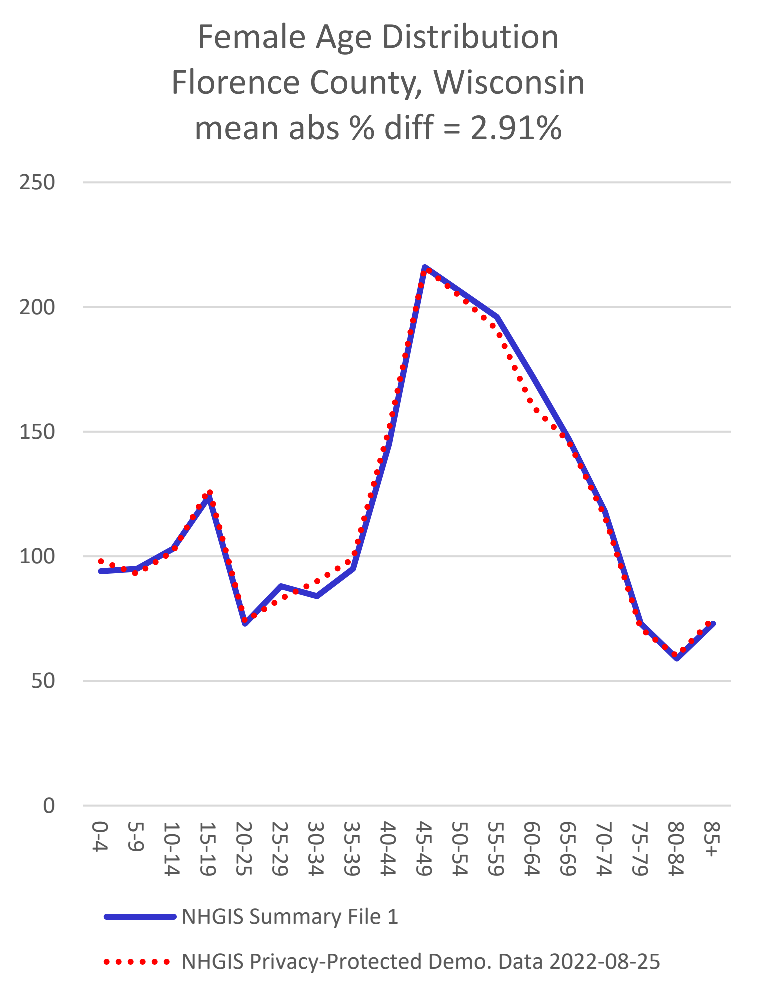


Females as % of Total Pop
Burnett County, Wisconsin
mean abs % diff = 1.37%

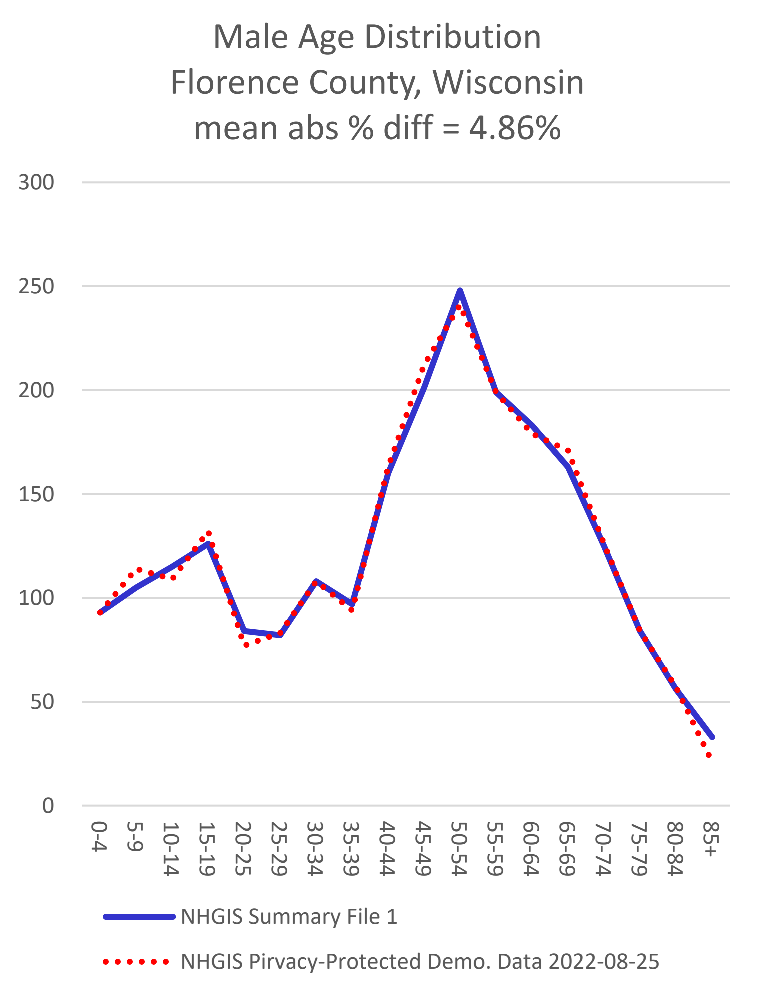


Race & Ethnicity
Burnett County, Wisconsin
mean abs % diff = 15.72%

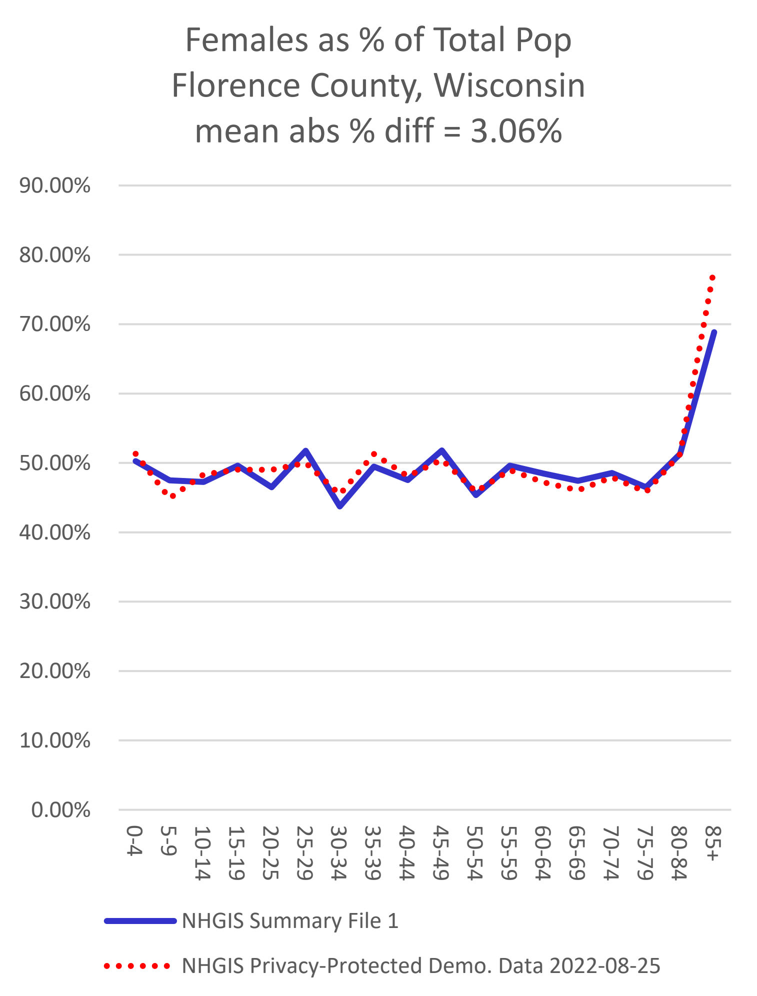




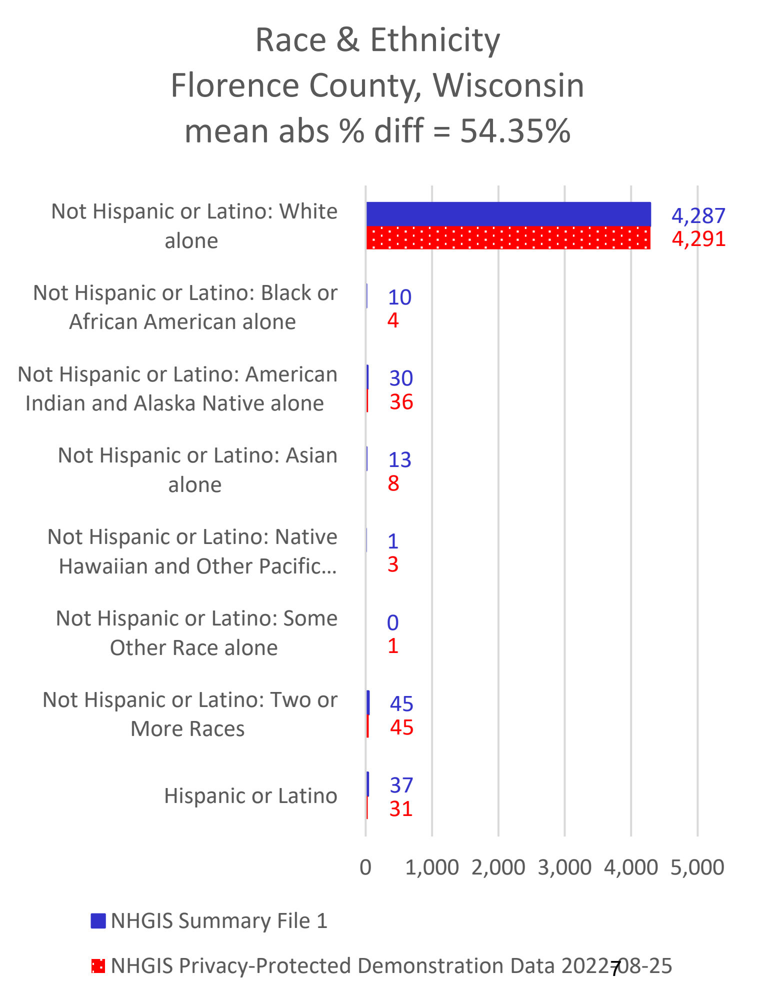
Female Age Distribution
Florence County, Wisconsin
mean abs % diff = 2.91%



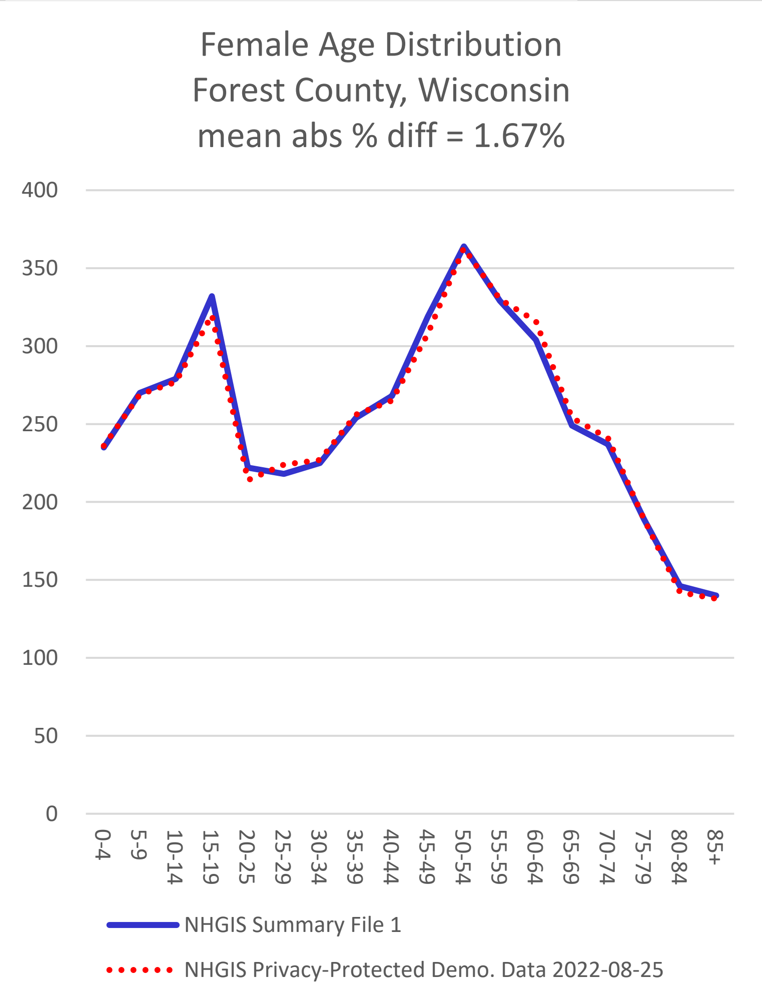
Male Age Distribution
Florence County, Wisconsin
mean abs % diff = 4.86%



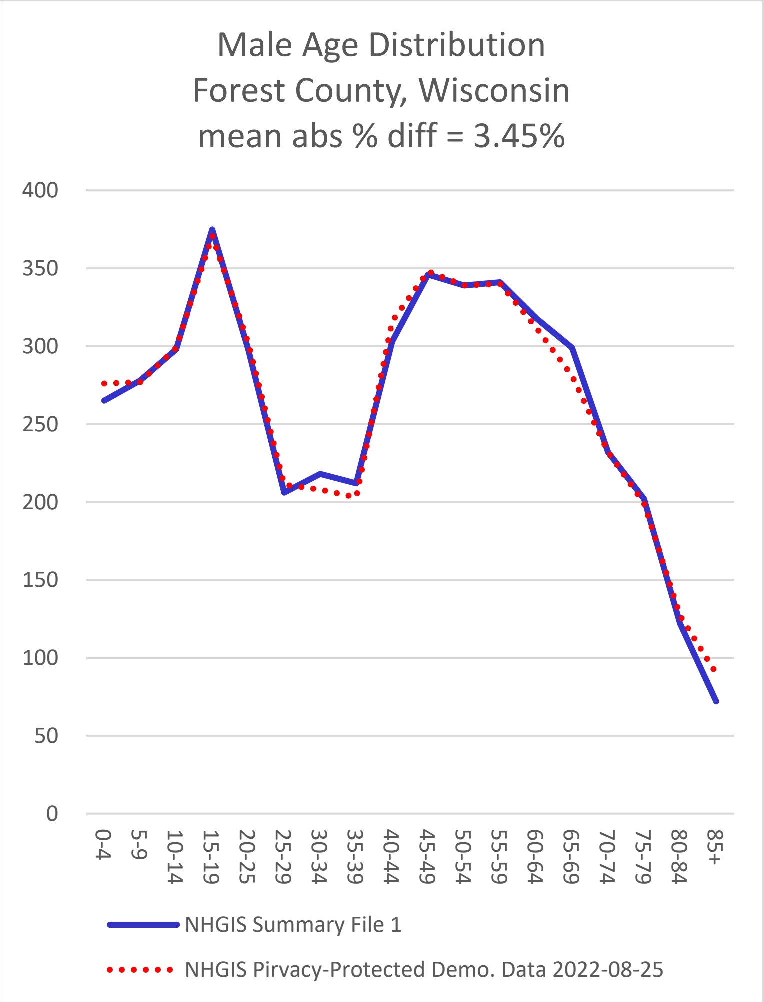
Females as % of Total Pop
Florence County, Wisconsin
mean abs % diff = 3.06%



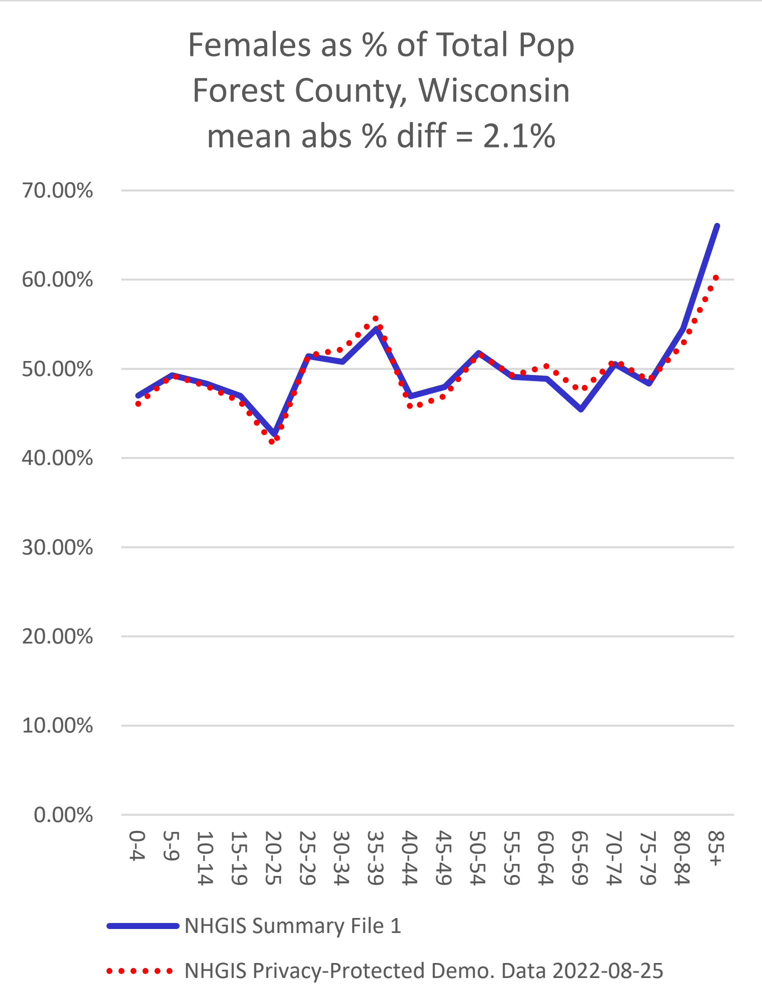
Race & Ethnicity
Florence County, Wisconsin
mean abs % diff = 54.35%



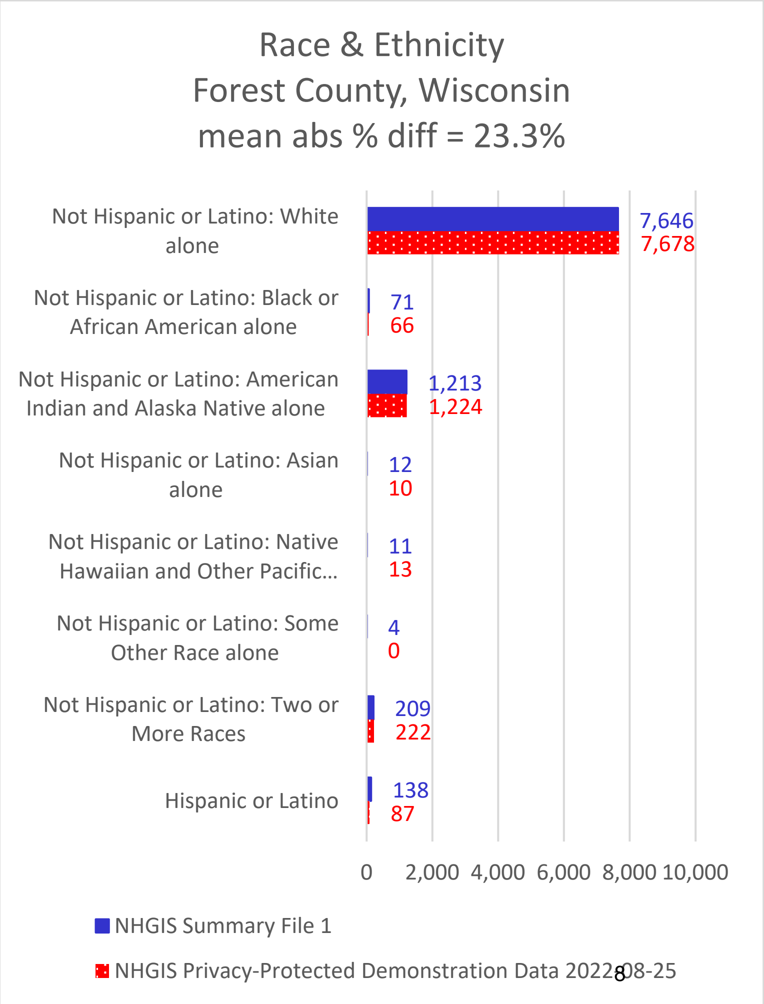
Female Age Distribution
Forest County, Wisconsin
mean abs % diff = 1.67%



Male Age Distribution
Forest County, Wisconsin
mean abs % diff = 3.45%

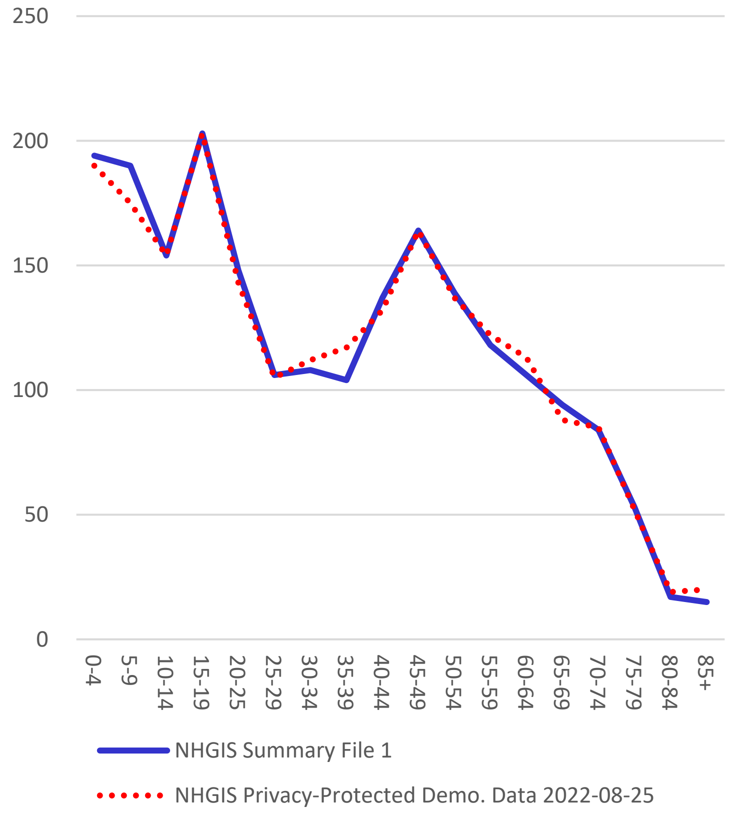


Females as % of Total Pop
Forest County, Wisconsin
mean abs % diff = 2.1%

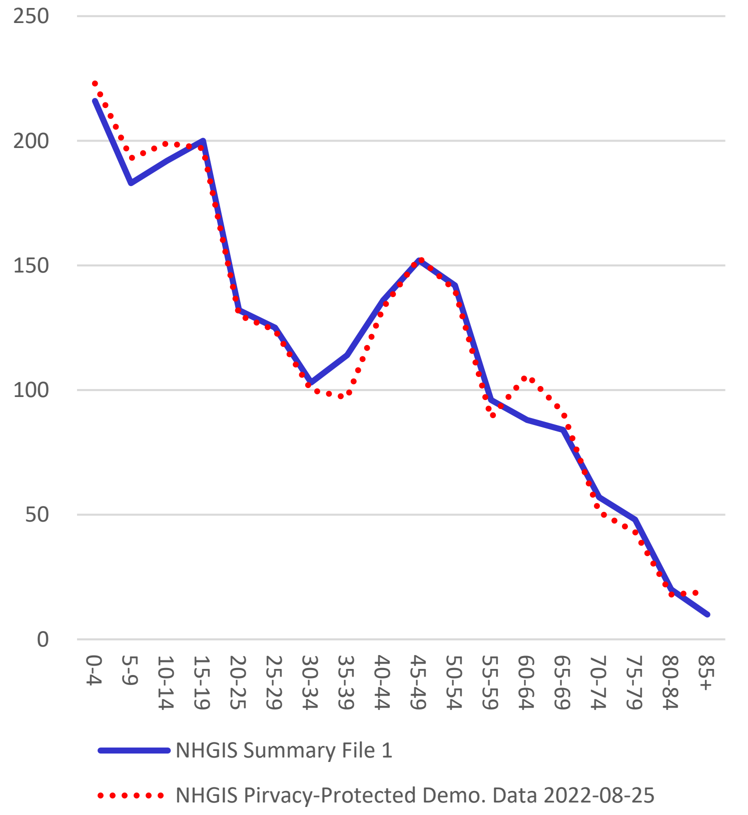


Race & Ethnicity
Forest County, Wisconsin
mean abs % diff = 23.3%

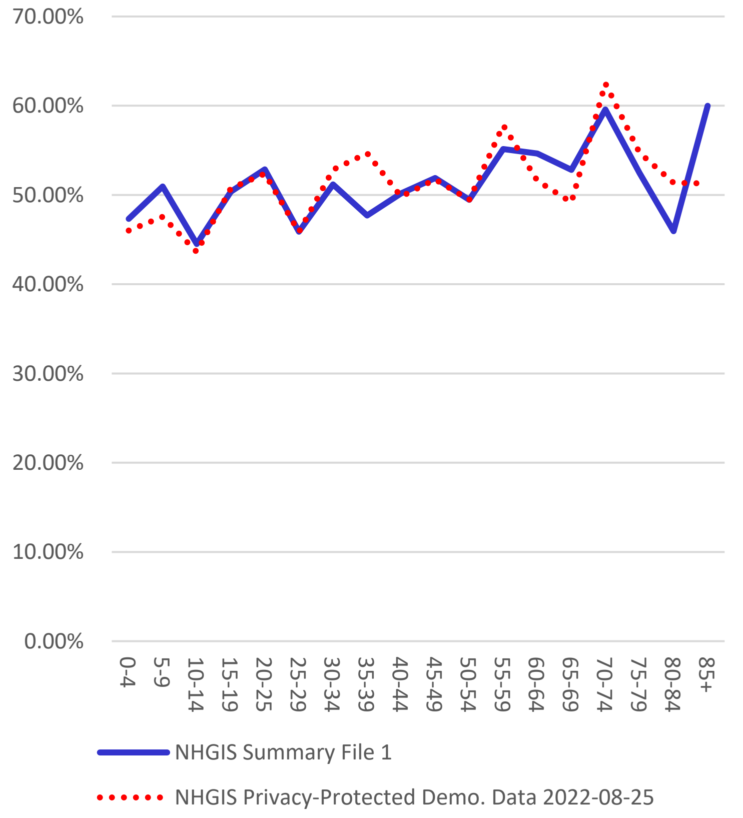
Female Age Distribution
Menominee County, Wisconsin
mean abs % diff = 5.56%



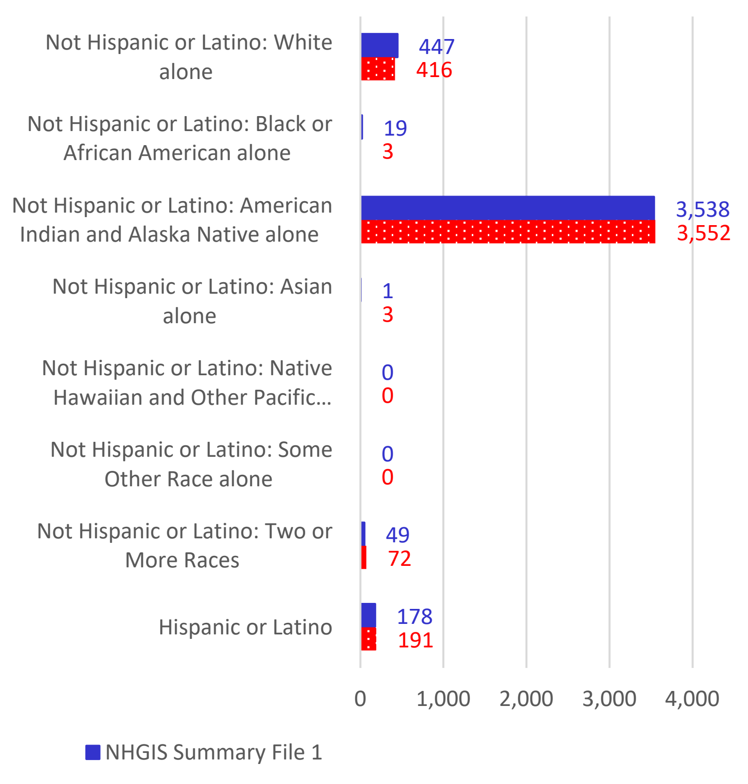
Male Age Distribution
Menominee County, Wisconsin
mean abs % diff = 10.85%

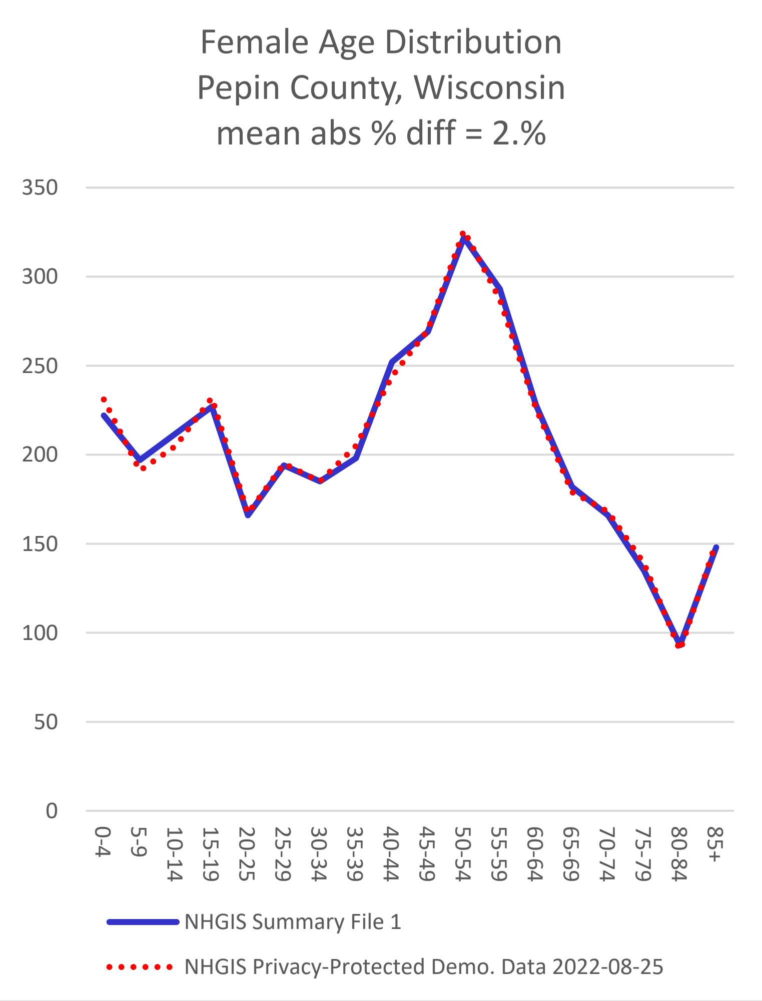


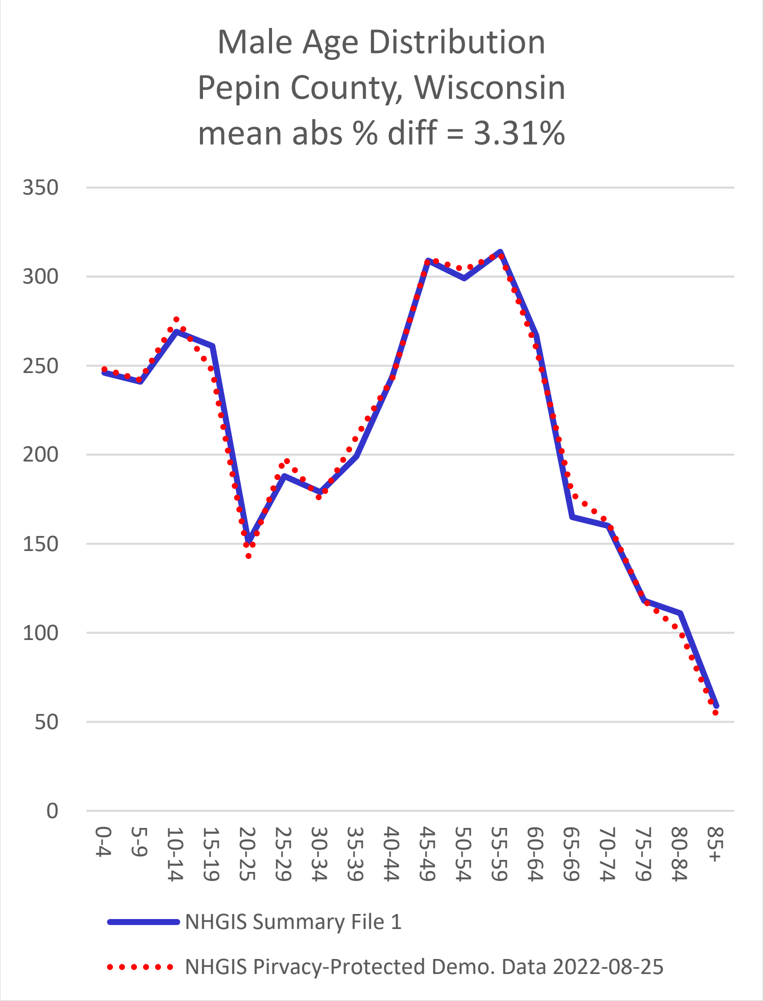
Females as % of Total Pop
Menominee County, Wisconsin
mean abs % diff = 4.71%

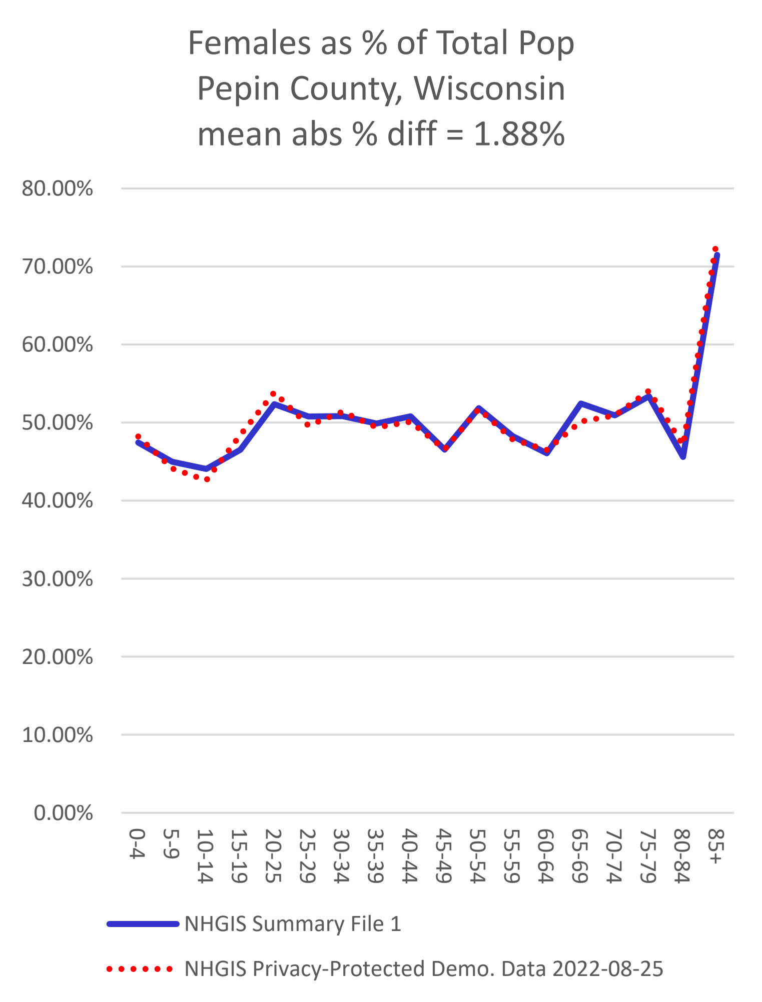


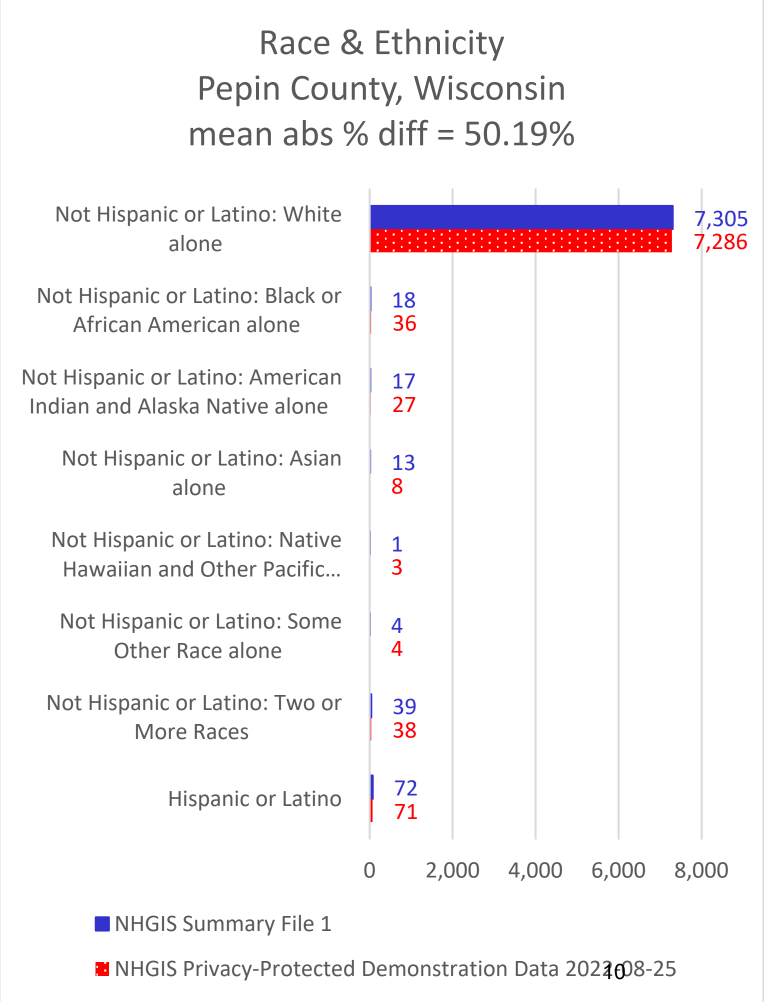
Race & Ethnicity
Menominee County, Wisconsin
mean abs % diff = 43.22%



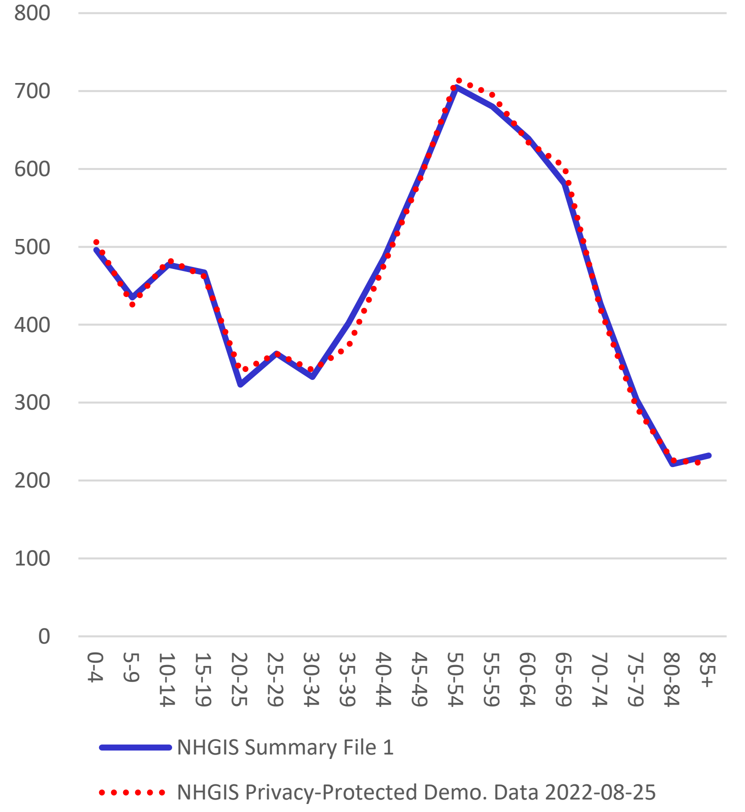

Female Age Distribution
Pepin County, Wisconsin
mean abs % diff = 2.%
350
300
250
200
150
100
50
0
0-4
5-9
10-14
15-19
20-25
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80-84
85+
NHGIS Summary File 1
NHGIS Privacy-Protected Demo. Data 2022-08-25


Male Age Distribution
Pepin County, Wisconsin
mean abs % diff = 3.31%
350
300
250
200
150
100
50
0
0-4
5-9
10-14
15-19
20-25
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80-84
85+
NHGIS Summary File 1
NHGIS Pirvacy-Protected Demo. Data 2022-08-25

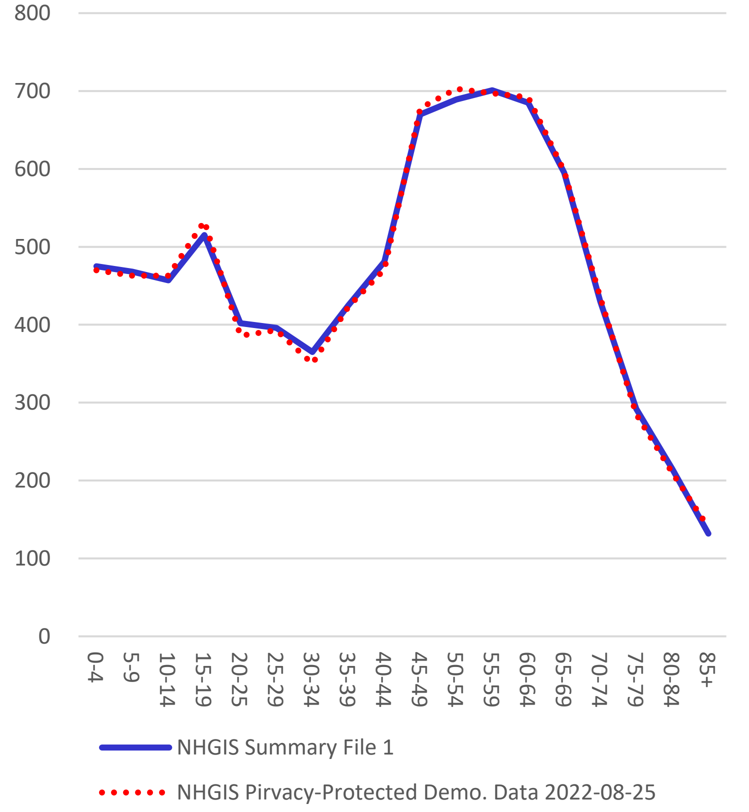

Females as % of Total Pop
Pepin County, Wisconsin
mean abs % diff = 1.88%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
0-4
5-9
10-14
15-19
20-25
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80-84
85+
NHGIS Summary File 1
NHGIS Privacy-Protected Demo. Data 2022-08-25


Race & Ethnicity
Pepin County, Wisconsin
mean abs % diff = 50.19%
Not Hispanic or Latino: White alone
7,305
7,286
Not Hispanic or Latino: Black or African American alone
18
36
Not Hispanic or Latino: American Indian and Alaska Native alone
17
27
Not Hispanic or Latino: Asian alone
13
8
Not Hispanic or Latino: Native Hawaiian and Other Pacific...
1
3
Not Hispanic or Latino: Some Other Race alone
4
4
Not Hispanic or Latino: Two or More Races
39
38
Hispanic or Latino
72
71
0
2,000
4,000
6,000
8,000
NHGIS Summary File 1
NHGIS Privacy-Protected Demonstration Data 2022-08-25

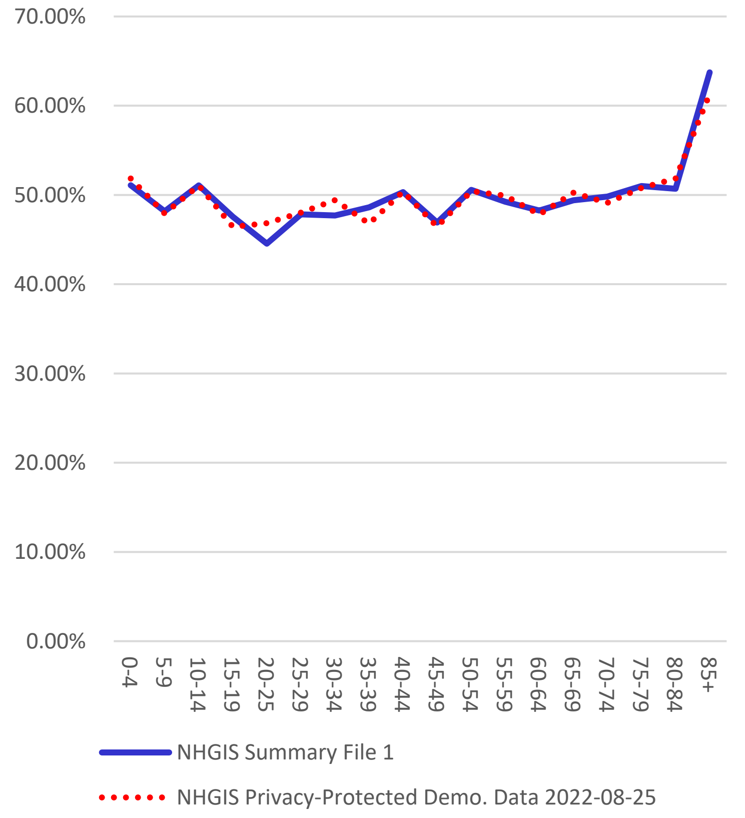
Female Age Distribution
Sawyer County, Wisconsin
mean abs % diff = 2.45%



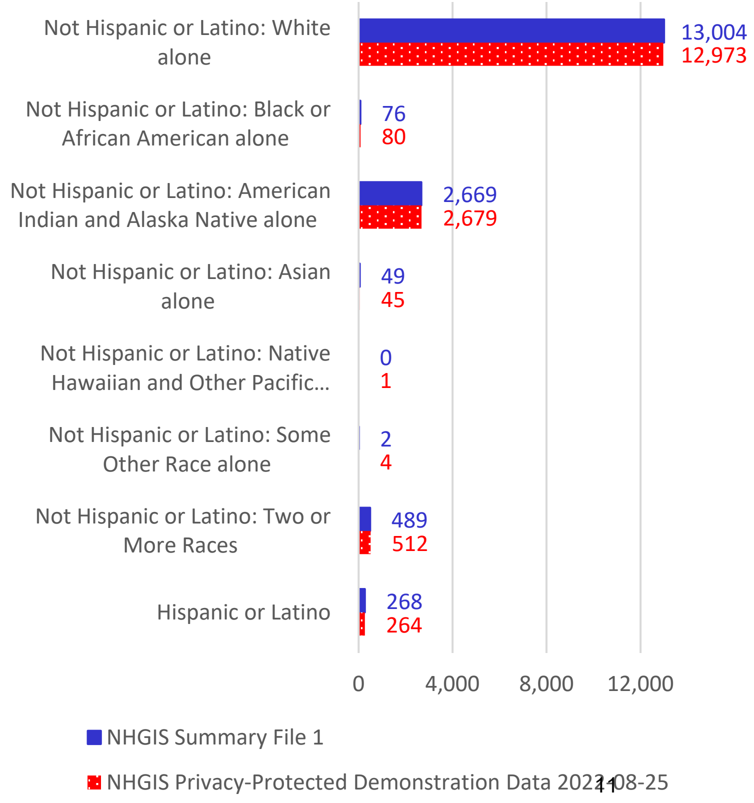
Male Age Distribution
Sawyer County, Wisconsin
mean abs % diff = 1.97%



Females as % of Total Pop
Sawyer County, Wisconsin
mean abs % diff = 1.7%



Race & Ethnicity
Sawyer County, Wisconsin
mean abs % diff = 27.53%



3. Bill OHare, O'Hare Data and Demographic Services, LLC

I am attaching a report that addresses the impact of DP on young children based on the August 25, 2022 DP Demonstration Product released by the Census Bureau.

I also want to take this opportunity to underscore one key point from my analysis. After DP is applied there are many Unified School Districts and Places with large errors for the population ages 0 to 4. I believe these large errors are the most serious problem caused by the application of DP and I urge the Census Bureau to try and reduce or eliminate these large errors.

Please let me know if you have any questions.

Bill O'Hare, PhD
President
O'Hare Data and Demographic Services, LLC

I am attaching a short paper reflecting my analysis of large errors for the population age 0-4 in the August 2022 DP demonstration product

Analysis of Census Bureau’s August 2022 Differential Privacy Demonstration Product: Implications for Data on Young Children

By
Dr. William P. O’Hare
September 2022

Analysis of Census Bureau’s August 25, 2022 Differential Privacy Demonstration Product: Implications for Data on Young Children

By

Dr. William P. O’Hare

Executive Summary

The U.S. Census Bureau is using a new method called differential privacy (DP) to help protect the confidentiality and privacy of respondents in the 2020 Census. This paper provides some information on how the use of DP in the 2020 Census is likely to impact the accuracy of data for young children (population ages 0 to 4).

The study is based on analysis of the most recent DP Demonstration Product released by the Census Bureau on August 25, 2022. The DP Demonstration Product issued on August 25, 2022 supersedes earlier DP Demonstration Products and focuses on data that will be in the 2020 Census Demographic and Housing Characteristics (DHC) file, which is scheduled to be released in May 2023.

The DHC file has most of the tables that were in Summary File 1 of the 2010 Census. The Demonstration Product released in August 2022 has data for population and housing units, but this analysis only examines data from the population file.

This paper presents analysis of the error introduced by DP by comparing the data as reported in the 2010 Census Summary File to the same data after the application of DP. According to the Census Bureau, the demonstration file released by the Census Bureau in August 2022 has been optimized for major use cases of the DHC tables.

Analysis presented in this paper found little impact of DP on data for young children for large (highly aggregated) geographic units like states or large counties.

However, the story is different for smaller geographic units. Many smaller areas have high levels of error in their data on young children after DP is applied. For example, the count of young children would exhibit absolute *percent* error of 5 percent or more in about 18 percent of Unified School Districts after DP is applied. The data also show that 64 percent of Unified School Districts had absolute *numeric* errors of 5 or more young children after DP is applied.

Errors of the magnitude shown above could have important implications for federal and state funding received by schools and for educational planning. Errors of this magnitude might impact formula funding that is based on Census-derived data and some schools would get less than they deserve.

Bigger absolute *percent* errors are evident for Hispanic, Black, and Asian young children in Unified School Districts. The mean absolute *percent* error for Non-Hispanic White young children was 5 percent compared to 28 percent of Hispanic young children, 35 percent for Black young children, and 45 percent for Asian young children. Differential accuracy among race and Hispanic Origin groups raises questions of data equity after DP is applied.

I also examined the accuracy/errors for the single year age 4 child population and found that errors for single year of age are particularly large. I found 52 percent of Unified School Districts had absolute *percent* errors of 5 percent or more for children age 4, and 59 percent had absolute *numeric* errors of 5 or more children age 4

The results are similar for Places. Analysis shows that 46 percent of Places (cities, village, and towns) had absolute *percent* errors of 5 percent or more for age 0 to 4, and 38 percent of Places had absolute *numeric* errors of 5 or more young children.

I believe the most important type of error introduced by the application of DP are the large errors introduced for some geographic units. Analysis shows that 2 percent of Unified School Districts have Absolute Percent errors of 25 percent or more. In terms of *numeric* errors, 5 percent of Unified School District have absolute *numeric* errors of 25 or more young children. I urge the Census Bureau to take steps to reduce or eliminate these large errors for I believe the large errors injected by DP that will be most problematic.

The application of DP also caused a number of impossible or improbable results. After the injection of DP in the 2010 Census data included in the August 2022 Census Bureau Demonstration Product (U.S. Census Bureau 2022d Table 18), there were 163,077 blocks nationwide (1.5 percent of all blocks) that had population ages 0 to 17, but no population ages 18 or over, compared to 82 such blocks before DP was applied This result has two important implications.

First, blocks with children and no adults are a highly implausible situation and the large number of such blocks may undermine confidence in the overall Census results.

Second, these implausible results are likely due to young children being separated from their parents in 2020 Census DHC processing with DP. This separation of children and parent in data processing is an ongoing concern for data on young children and the production of future tables for children. This issue is particularly important in introducing DP into the American Community Survey, which is a key source of child well-being measures (O'Hare 2022b). To understand the well-being of children, it is critical to understand the situation of a child's parents or caretakers. Moreover, if the same separation of children from their parents and caregivers occurs in the application of DP to the American Community Survey, it will eliminate reliable child poverty data which is based on household income. Child poverty rates are one of the most important measures of child well-being.

Based on the errors for the young child population with the privacy parameters for DP used in the August 2022 DP Demonstration Product, and the lack of clarity about the level of privacy protection from DP, I recommend the Census Bureau take steps to reduce the size of errors injected into the 2020 Census DHC file and in particular focus on trimming or eliminating the number of large errors.

This paper is meant to provide stakeholders and child advocates with some fundamental information about the level of errors DP is likely to inject into the 2020 Census data for the population ages 0 to 4. There are a couple of reasons for sharing this information with child advocates now. The 2020 Census results for some localities may include situations where the number of young children reported looks suspect. It is important to make sure child advocates are aware of the potential impact of DP so they can explain odd child statistics to local leaders.

There is a second reason for sharing this information with state and local child advocates. The U.S. Census Bureau is looking for feedback on the use of DP in the 2020 Census. The Census Bureau is looking for cases where census data are used to make decisions and the Census Bureau is asking data users to examine the DP

Demonstration Product to see if the error injected by DP make the data unfit for use. After reading this report, I hope you will convey your thoughts to the Census Bureau.

There is some latitude in how much error the Census Bureau will inject into the DHC files so feedback from census data users is important. If many users feel the current level of precision for data on young children in DP Demonstration Product is not accurate enough for some uses, there is a chance the Census Bureau could make the final data more accurate.

Stakeholders, child advocates, and data users should take advantage of this opportunity to communicate their thoughts to the Census Bureau before Census Bureau’s Data Stewardship Advisory Committee makes a final decision on the privacy parameters to be used in the DHC file when it is released in May of 2023. Comments on the implications of DP in the August 2022 Demonstration File are due September 26, 2022, **Comments and responses can be sent to 2020DAS@census.gov.**

Analysis of Census Bureau's August 25, 2022 Differential Privacy Demonstration Product: Implications for Data on Young children

By
Dr. William P. O'Hare

Introduction

The U.S. Census Bureau is using a new method called differential privacy (DP) to help protect confidentiality and privacy of Census respondents in releasing data from the 2020 Census.[1] Analysis in this paper uses several measures to assess the accuracy of census data for young children after DP is applied. Young children are defined in this report as those ages 0 to 4. The analysis is based on the Demonstration Product released on August 25, 2022, which is the most recent available from the Census Bureau. This is the last Demographic and Housing Characteristics (DHC) Demonstration Product file the Census Bureau will release before they determine the final production parameters for the DHC file to be released in May 2023.

In short, DP injects errors in the data provided by respondents to make it more difficult for someone to be identified in the Census records. Adding or subtracting random numbers to the census results makes it more difficult to identify data for specific respondents because the data in the published census results no longer match what respondents submitted. The U.S. Census Bureau (2020e) provides more information

[1] The terminology in this arena can be confusing. Differential Privacy is sometimes called "formal privacy." The system developed for the 2020 Census DHC file has also been called the Top Down Algorithm or TDA. Since the application of differential privacy occurs within the Census Bureau's Disclosure Avoidance Systems (DAS) that term has sometimes been used to describe the use of differential privacy. To avoid confusion, I use the term differential privacy (DP) here to distinguish the version of DAS that includes DP from other versions of DAS.

on the use of DP in the 2020 Census along with regular updates of their work (U.S. Census Bureau 2020c). In the fall of 2021, the Census Bureau released a primer on DP. (U.S. Census Bureau 2021d).

For an independent look at differential privacy see Boyd (2019) or Bouk and Boyd (2021). Hotz and Salvo (2020) offer a good review of DP early in the Census Bureau's development. A good overview of the evolution of the DP issue at the Census Bureau is provided by Boyd and Sarathy (2022).

It is fair to say that the introduction of DP in the 2020 Census has become a very controversial issue. In their review of the development of the DP issue over the past few years, Boyd and Sarathy (2022, page 1) conclude, "When the U.S. Census Bureau announced its intention to modernize its disclosure avoidance procedures for the 2020 Census, it sparked a controversy that is still underway."

One reason to focus on the impact of DP on the population ages 0 to 4 is the high net undercount of that population in the Census. Results of the 2020 Census evaluation using the Demographic Analysis method, show a net undercount of 5.4 percent for young children which was much higher than any other age group (U.S. Census Bureau 2022c).

Recent trends are also unsettling. From 1950 to 1980, the young children and adults had similar decade-to-decade improvement in terms of census coverage. However, after 1980 the trajectories were quite different. The coverage for adults continued to improve while the coverage of young children decreased dramatically (O'Hare 2022a). The net undercount of young children in the 2020 Census (5.4 percent) is higher than the young children undercount in the 1950 Census. I am not aware of any other population group where census coverage is worse in the 2020 census than it was in the 1950 Census.

There are a couple of perspectives one could take regarding the high net undercount of young children and DP. On one hand, since the 2020 Census data for young children already has more error than data for other age groups, perhaps the amount of error injected by DP should be limited for this group. It does not seem fair to inject more error into data for groups that already have a lot of error in their census results. On the other hand, one might think that since the 2020 Census data for young children already has a lot of error, the added error from DP will not make much difference.

I focus first on data accuracy for Unified School Districts because schools are the public institution most closely associated with the child population and schools use demographics in a variety of ways. I next look at data for Places. Places include big cities and small villages. They typically have policymaking authority, and they often provide programs for young children such as childcare or preschool programs.

Several issues regarding DP are addressed in the Discussion section including the high error rate for blocks, breaking the relationship between children and parents, questions of equity, and the extent to which DP contributes to the lack of public trust in the census.

Background on Privacy in the Census

In every census, the U.S. Census Bureau faces a trade-off between privacy protection and accuracy. According to the U.S. Census Bureau (2020d),

> “One of the most important roles those national statistical offices (NSOs) play is to carry out a national population and housing census. In so doing, NSOs have two data stewardship mandates that can be in direct opposition. Good data stewardship involves both safeguarding the privacy of the respondents who have entrusted their information to the NSOs as well as disseminating accurate and useful census data to the public.”

The problem that DP is designed to fix is complicated as is the implementation of DP. The passage below from the U.S. General Accountability Office (2020, page 14) is the best short description I have seen on this issue.

> “Differential privacy is a disclosure avoidance technique aimed at limiting statistical disclosure and controlling privacy risk. According to the Bureau, differential privacy provides a way for the Bureau to quantify the level of acceptable privacy risk and mitigate the risk that individuals can be reidentified using the Bureau’s data. Reidentification can occur when public data are linked to other external data sources. According to the Bureau, using differential privacy means that publicly available data will include some statistical noise, or data inaccuracies, to protect the privacy of individuals. Differential privacy provides algorithms that allow policy makers to decide the trade-offs between data accuracy and privacy. “

It is important to note that the U.S. Census Bureau has used methods to help avoid disclosure of individual census respondents for many decades. According to U.S. U.S. Census Bureau (2018) some method of disclosure avoidance has been used by the U.S. Census Bureau since 1970. The 2010 Census data include some changes to original responses to help avoid disclosure of information about individual respondents, largely using a method called swapping.

The application of differential privacy allows the Census Bureau to control the amount of error injected into the data which is largely controlled by a parameter called “Epsilon.” A higher-level of Epsilon means less error and more risk of violating confidentiality and a lower level of Epsilon means more error and less risk of violating confidentiality. In the latest material from the Census Bureau, Epsilon has been replaced with a term called Rho. It is my understanding Rho works the same way as

Epsilon in that a higher value means more accuracy and a lower value means more privacy protection. The point here is that the Census Bureau has control over how much error to inject into the data.

Measuring Accuracy

There is no consensus on exactly what measures should be used to assess the accuracy of DP-infused data, and there is no single benchmark to determine if DP-infused figures are "accurate enough for use." The U.S. Census Bureau (2020a) has suggested several measures of accuracy that could be used to evaluate the DP-infused data.

Like the Census Bureau's assessment of DP-infused data, I provide data for both absolute *numeric*al errors and absolute *percent* errors because either can be important and using both perspectives provide a more complete picture of the error profiles for geographic units. It may be a bit confusing presenting both *numerical* and *percent* errors, so I italicize the terms for help readers more easily distinguish which measure is being discussed.

For simplicity I only look at a few key measures here, but they provide sufficient information to reach some conclusions. The measures used here (mean absolute *numeric* error, mean absolute *percent* error, and large errors) are a subset of those discussed by the Census Bureau.

The DP demonstration file released by the Census Bureau on August 25, 2022, provides DP-infused data from the 2010 Census which can be compared to the 2010 Census data without DP to understand the likely impact DP has on data accuracy.

Errors are defined here as the difference between the data as originally reported in the 2010 Census Summary File and the same data after DP has been injected. The data from the Summary File is sometimes referred to as data without the application of DP in this report. Specifically, I subtract the value of the data with DP from the corresponding data without DP (Summary File) to find the error. For percentages, the difference is divided by the data without DP (i.e., Summary File) value.

I include a measure the Census Bureau calls the Mean Absolute Error (I label this Mean Absolute *Numerical* Error in the tables to distinguish it from the Mean Absolute *Percent* Error) and I also include the Mean Absolute *Percent* Error.

An absolute error reflects the magnitude of the error regardless of direction. A geographic unit with an absolute error of 10 percent could be 10 percent too high or 10 percent too low. Absolute errors are used to make sure positive errors and negative errors do not cancel each other out and make it appear as if there are no errors.

Percent error reflects the size of the error relative to the size of the population. An error of a given magnitude (say 10 young children) may be trivial in large Places but very significant in smaller Places. For example, a numeric error of 10 young children in a school district of 1,000 young children is only a 1 percent error, but a *numeric* error of 10 young children in a school district of 100 is a 10 percent error.

In addition to measures of average error, I include analysis on the number and percent of geographic units that have relatively large errors. I use two sets of benchmarks to identify large errors: one for absolute *numeric* errors and one for absolute *percent* errors.

The number and percent of large errors are likely to be the most important measures of accuracy in the 2020 Census. Large errors are likely to be a statistical problem and a public relationship problem for the Census Bureau, particularly if the errors are accompanied by large swings in funding. Data from the Census is often used to distribute federal and state dollars based on population (O'Hare 2020a: Reamer 2020, O'Hare and Rashid 2022: The Annie E. Casey Foundation, 2018). Large errors can result in implausible or impossible results. Such results are likely to cast suspicion on all the data from the Census Bureau and it is likely to undermine the confidence people have in all the census data.

Data Used in This Study

The Demonstration Product released in August 2022 reflects ongoing work at the Census Bureau. Starting in October 2019, the Census Bureau has released several Demonstration Products that reflect the injection of DP into 2010 Census data. The first official data from the 2020 Census with DP infused was the redistricting data file released by the Census Bureau in August 2021.

The DP Demonstration Product examined here is related to the Demographic and Housing Characteristics file that is scheduled to be released in May 2023. The Census Bureau (U.S. Census Bureau 2022d) has provided some measures of accuracy for the DP Demonstration Product, but they are somewhat limited.

Related to previous DP releases, my analysis of the DHC DP Demonstration Product released in March 2022, is available on the Count All Kids website (O'Hare 2022c). The Census Bureau's summary of all comments submitted in relation to the

March 2022 DP Demonstration Product are also available (U.S. Census Bureau 2022f).

The data used in my analysis were originally provided by the Census Bureau. The IPUMS- NHGIS unit at the University of Minnesota processed the Census Bureau files and put the data into more user-friendly tables. I analyzed the data produced by IPUMS-NHGIS unit which are available at https://nhgis.org/privacy-protected-demonstration-data

Geographic units where there were zero people ages 0 to 4 in either the 2010 data with DP or without DP were removed from the files for analysis. Observations with zeros for key measures produce very unusual results. This analysis does not include data for Puerto Rico.

Results for Age 0 to 4 in Four Kinds of Geographic Units

Table 1 provides a few key accuracy measures for the population ages 0 to 4 for four kinds of geographic units. These units were selected because they all have significant policy-making power regarding programs for children and they range widely in terms of population size.

The results shown in Table 1 indicate that DP is unlikely to have much of an impact on the young child data for states. The mean absolute *numeric* error for states for the population ages 0 to 4 is about 100 young children and the mean absolute *percen*t error rounds to zero.

Also, DP is unlikely to have much impact on young child county data for most counties. The mean absolute *numeric* error for counties is about 8 young children and mean absolute *percent* error is 0.92.

However, of the 3,142 counties examined here 36 percent (1,130) had less than 1,000 children ages 0 to 4 based on the Summary File results. For this subset of counties, DP may distort the data to a considerable degree. For the 1,130 counties with less than 1,000 young children, the mean absolute *numeric* error for ages 0 to 4 was 6 and the mean absolute *percent* error was 2.1.

Table 1 Key Statistics for Absolute *Numeric* and Absolute *Percent* Errors* for Children Ages 0 to 4 for Selected Geographic Units

	States	Counties***	Unified School Districts****	Places *****
Number of Units in the Analysis	50	3,141	11	28,548
Mean Size of District (Children ages 0-4 based on Summary File)	403,375	6,429	1,860	545
Mean Absolute *Numeric* Error**	100	8	9	5
Mean Absolute *Percent* Error	rounds to zero	0.92	4	13
Percent of Units with Absolute *Numeric* Errors of 5 or more Children	98	62	64	38
Percent of Units with Absolute *Percent* Errors of 5% or more****	0	3	18	46

Source: Author's analysis of Demonstration Product data released by the Census Bureau on August 25, 2022. Data from IPUMS NHGIS, University of Minnesota www.nhgis.org

Data in this table does not include Puerto Rico or geographic units with zero population age 0 to 4 in 2010 Summary File

* in this paper errors reflect the difference between the 2010 Census data without and with DP injected (SF- DP).

** The Census Bureau calls this measure Mean Absolute Error. I include the word "Numeric" to distinguish it from Mean Absolute Percent Error.

***DC is not included in the state data but is included in the county data

**** based on percentages rounded to two decimal points

***** this include both incoporated places and Census Designated places

The situation is different for Unified School Districts and Places (shown in Table 1), where DP is likely to cause larger distortions (percentage-wise) for the young child population. The mean absolute *numeric* error for Unified School Districts is 9 young children and it is 5 young children for Places. The mean absolute *percent* error for United School Districts is 4 percent and it is 13 percent for Places.

In my opinion the bigger problem is the number of extreme errors for these geographic units. For Unified School Districts and Places, the share of units that have extreme errors is substantial. Table 1 shows that 64 percent of Unified School District have absolute *numeric* errors of 5 or more children and 18 percent have absolute *percent* errors of 5 percent or more. For Places, 38 percent have absolute *numeric* errors of 5 or more children, and 46 percent have absolute *percent* errors of 5 percent or more. These extreme errors are more consequential than the mean figures. Accuracy for Unified School Districts and Places are explored in more detail in the next two sections of this report including more information on extreme errors.

Application of Differential Privacy to School District Data

The analysis first focuses on Unified School Districts since schools are the largest public institution focused on children. The Census Bureau reports there were 61.6 million children ages 3 to 17 enrolled in schools in 2019 (U.S. Census Bureau 2021a).

Schools often provide preschool programs for those under age 5. The Census Bureau shows there were over 5 million children enrolled in preschool in 2019, and more than half of all children age 3 and 4 are in preschool or nursey school (McElrath et al. 2022)

Reamer (2020) shows that $39 billion of federal funds were distributed by the U.S. Department of Education to states and localities in FY 2017 based on census-derived data. Table 2 shows programs run by the U.S. Department of Education that distribute federal funds to state and localities based on census-derived data. In addition, many other government programs also use census-derived data to distribute funds targeted to children. This underscores why the accuracy of the population figures from the Census are so important.

Overall, Reamer (2020) identified 316 federal programs that use census-derived data to distribute about $1.5 trillion to states and localities in Fiscal Year 2017. About two-thirds of the 315 programs use substate data which underscores the importance of small area census data. When one is talking about billions of dollars, a small percent error can translate into a large dollar amount.

Table 2. Federal Programs in the U.S. Department of Education that Distribute Funds to States and Localities based on Census-derived Data

	Amount Distributed in FY 2017
Adult Education - Basic Grants to States	$581,955,000
Title I Grants to LEAs	$15,459,802,000
Special Education Grants	$12,002,848,000
Career and Technical Education - Basic Grants to States	$1,099,381,000
Vocational Rehabilitation Grants to the States	$3,121,054,000
Rehabilitation Services - Client Assistance Program	$13,000,000
Special Education - Preschool Grants	$368,238,000
Rehabilitation Services - Independent Living Services for Older Individuals Who are Blind	$33,317,000
Special Education-Grants for Infants and Families	$458,556,000
School Safety National Activities	$68,000,000
Supported Employment Services for Individuals with the Most Significant Disabilities	$27,548,000
Program of Protection and Advocacy of Individual Rights	$17,650,000
Twenty-First Century Community Learning Centers	$1,179,756,000
Gaining Early Awareness and Readiness for Undergraduate Programs	$338,831,000
Teacher Quality Partnership Grants	$43,092,000
Rural Education	$175,840,000
English Language Acquisition State Grants	$684,469,000
Supporting Effective Instruction State Grants	$2,055,830,000
Grants for State Assessments and Related Activities	$369,051,000
Teacher Education Assistance for College and Higher Education Grants	$90,955,000
Preschool Development Grants	$250,000,000
Student Support and Academic Enrichment Program	$392,000,000
Total	$38,831,173,000

Source: Counting for Dollars. https://gwipp.gwu.edu/counting-dollars-2020-role-decennial-census-geographic-distribution-federal-funds

It is also clear that census-related data are often used by states to distribute state government money, but as far as I can tell, there is no systematic data on how much money is distributed by states based on Census data (O'Hare 2020a).

At the National Academy of Sciences, Committee on National Statistics workshop on DP which was held in December 2019 there were several presentations reflecting implications of DP-infused data for young children and school districts (Vink 2019; O'Hare 2019; Nagle and Kuhn 2019). Note that some of these analyses are now outdated but they may be useful for framing issues.

O'Hare (2021) focuses on the accuracy of population ages 0 to 17 for Unified School Districts based on data from the Census Bureau's redistricting file released in August 2021. In addition, my analysis of the DP Demonstration Product's impact on data for young children based on the DP file released in March 2022 by the Census Bureau can be found on the Count All Kids website (O'Hare 2022c).

Demographic data are used for several important school district applications. Population projections are often used to plan for expanding (or reducing) school facilities, staff, and other school-related needs. Demographic projections are typically based on Decennial Census data.   Current and projected demographic data are often used to construct attendance boundaries to keep classrooms from becoming overcrowded. Constructing attendance boundaries often include sensitivity to racial composition, so small area demographics by race are important. Such activities often require very small area data such as census blocks. Demographers who work extensively with school districts report that census blocks are a critical geographic unit for their work (Cropper et al. 2021).

Many school districts are governed by school boards which are often elected from single member districts.   Such districts must meet the usual legal requirements of redistricting such as having districts with equal population size. Such redistricting must also meet the requirements of the Voting Rights Act, which means small area tabulations of population by race and Hispanic origin are important.

Once children get into the K-12 school system, school systems have pretty good data for forecasting the number of children to expect in each grade the following year. From that perspective it is the cohort age 0 to 4 that is the biggest unknown for many school systems. Therefore, this is the most important age group for examining the amount of error injected by DP.

DP has a bigger impact, percentage-wise, in smaller populations and the majority of Unified School Districts are relatively small.   Many of the 10,864 Unified School Districts in this analysis are very small; 7,475 (69 percent of all Unified School Districts) had a young child population of less than 1,000, and 1,454 districts (13 percent of all districts) had a young child population less than 100 in the 2010 Census. The translation of small *numeric* errors into large *percent* errors is also more apparent in looking at data for Hispanic, Black, and Asian groups within Unified School Districts because those are typically smaller population groups.

Table 3 shows several measures of accuracy/error for 10,864 Unified School Districts in the 2010 Census used in this analysis.[2] The data are provided for all young children (all races) as well as for Non-Hispanic White Alone young children, Hispanic young children, Black Alone young children, and Asian Alone young children. For the remainder of this report when I use the term Black or Asian, it means Black alone or Asian alone. Other race groups were not examined here because the numbers were small, they were often highly clustered, and time was limited.

Data in Table 3 show the majority of Unified School Districts have at least one Black child, one Hispanic child, and one Asian child. But many districts have relatively few young children of color. The average number of Hispanic young children in Unified School Districts where there was at least one Hispanic was 521, for Blacks it was 384. and for Asians it was 143. These numbers are well below the overall average of 1,860 young children. The relatively small number of Black, Hispanic, and Asian young children in many districts results in these groups having larger absolute *percent* errors.

Table 3 shows the mean absolute *numeric* error for all young children (all races) in Unified School Districts is 9 young children. Data in Table 3 shows for all children, the mean absolute *percent* error was 4. But these measures mask big differences among race and ethnic groups.

The mean absolute *numeric* errors for race and Hispanic Origin groups are smaller than for all children (8 for Hispanic young children 6 for Black young children,

[2] Recall that districts where there was a zero for population age 0 to 4 in the DP or SF file were not included in the analysis. Also, recall Puerto Rico is not included.

and 4 for Asian young children), compared to 9 for all children, as these are smaller population groups in general

On the other hand, mean absolute *percent* error was 4 percent for all children, 28 percent for Hispanic, 35 percent for Blacks young children, and 45 percent for Asian young children (Table 3).

Table 3 Summary of Key Error* Statistics for Children Ages 0 to 4 for Unified School Districts by Race and Hispanic Origin

	All young children	Non-Hispanic White Alone	Hispanic	Black**	Asian**
Number of units in the analysis	10,864	10,841	10,238	7,548	6,251
Mean number of young children in district (in group column heading)	1,860	945	521	384	143
Mean absolute numeric error***	9	9	8	6	4
Mean absolute percent error	4	5	28	35	45
Percent of units with errors of 5 or more young children	64%	64%	49%	40%	31%
Percent of units with errors of 5% or more	18%	23%	66%	63%	70%

Source: Author's analysis of Demonstration Product data released by the Census Bureau on August 25, 2022 after being processed by IPUMS NHGIS at the University of Minnesota www.nhgis.org

Data in this table does not include Puerto Rico or geographic units with zero population age 0 to 4 in 2010 Summary File or DP-infused file.

* in this paper errors reflect the difference between the 2010 Census data without and with DP injected.

** The Census Bureau calls this measure Mean Absolute Error. I include the word "Numeric" to distinguish it from Mean

Recall that absolute errors reflect the magnitude of the error without regard to the direction of the error. Absolute errors are used in this analysis so that positive and negative errors do not cancel each other out in constructing an average or mean.

<u>Large Errors in Unified School Districts</u>

Means or averages are helpful, but they do not reveal the full story. Large errors can be problematic even if the overall mean error is relatively low. An examination of the distribution of Unified School Districts by error size can provide more information on the relative accuracy of the DP-infused data.

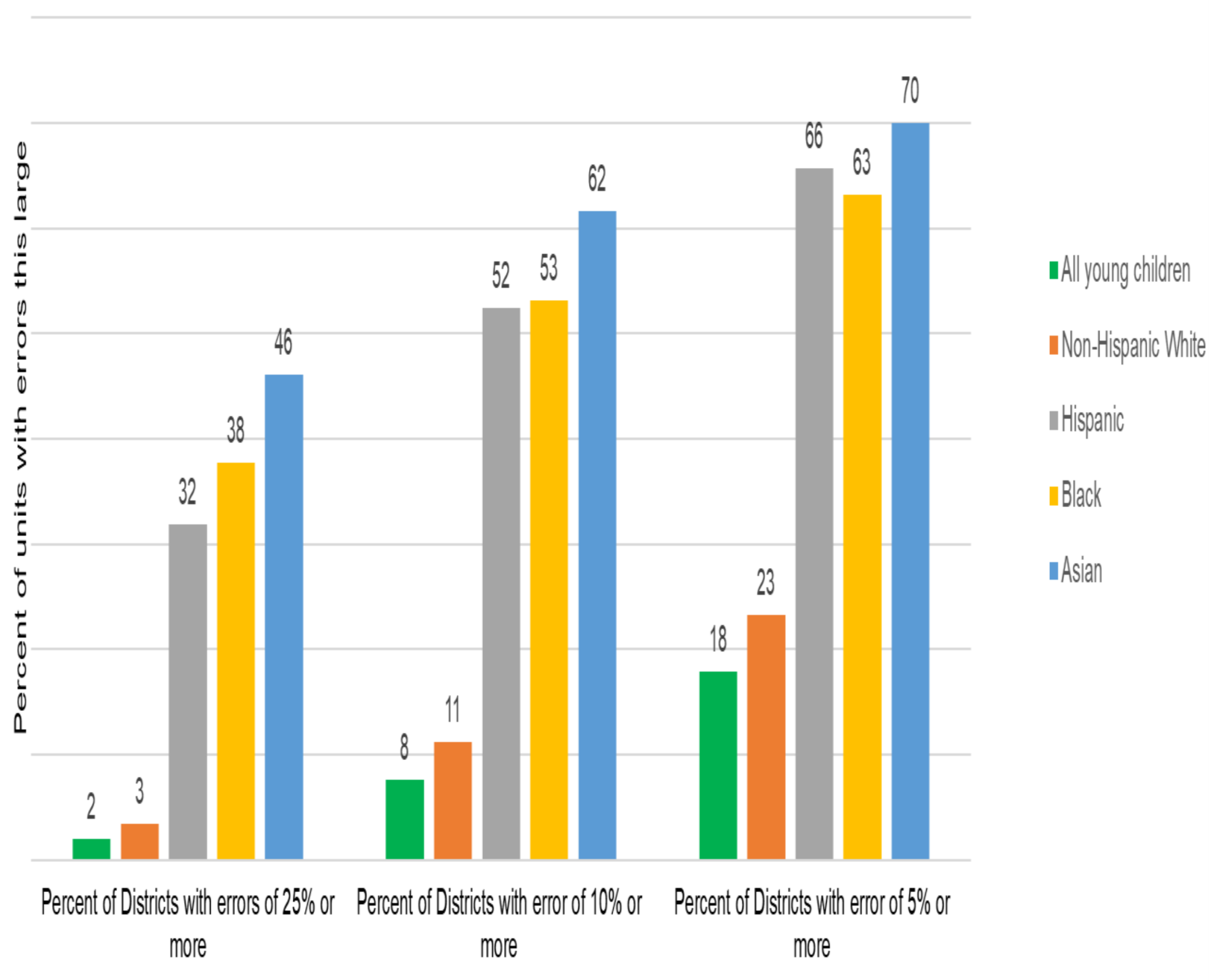
There is no consensus on what constitutes a large error and definitions probably vary with different applications. I show three benchmarks for large absolute *percent* errors. The 5 percent or more and 10 percent or more categories are used in several publications. I added the 25 percent plus category to look at the most extreme errors. Errors of 25 percent or more are likely to be very problematic. These thresholds are judgmental, but they provide a reasonable range of errors.

To be clear, the districts with more than 25 percent with large errors are also counted in the categories for more than 10 percent error and more than 5 percent error.

Distributions of absolute *percent* errors are shown in Figure 1 which shows that for all young children, 18 percent of districts had absolute *percent* errors of 5 percent or more, compared to 23 percent of Non-Hispanic White Alone, 66 percent for Hispanic young children, 63 percent for Black young children, and 70 percent for Asian young children. Since minority groups are smaller in population size, it is not surprising that there are more extreme absolute *percent* errors. There is a similar pattern by race and Hispanic Origin for other benchmarks.

In the largest error category (25 percent or more) the numbers are quite low for all young children and non-Hispanic whites alone young children, but quite high for Black, Hispanic, and Asian young children. Figure 1 shows that 32 percent of Unified School Districts have absolute *percent* errors of 25 percent or more for Hispanics, compared to 38 percent for Blacks and 46 percent for Asians. Figure 1 also shows that for young children of color, absolute *percent* errors of 25 percent or more are not unusual. Only two percent of Unified School Districts have absolute *Percent* Errors of 25 percent or more, but this amount to about more than 200 Districts nationwide.

Figure 1. Distribution of Absolute Percent Errors for Population Ages 0 to 4 for Unified School Districts by Race and Hispanic Origin

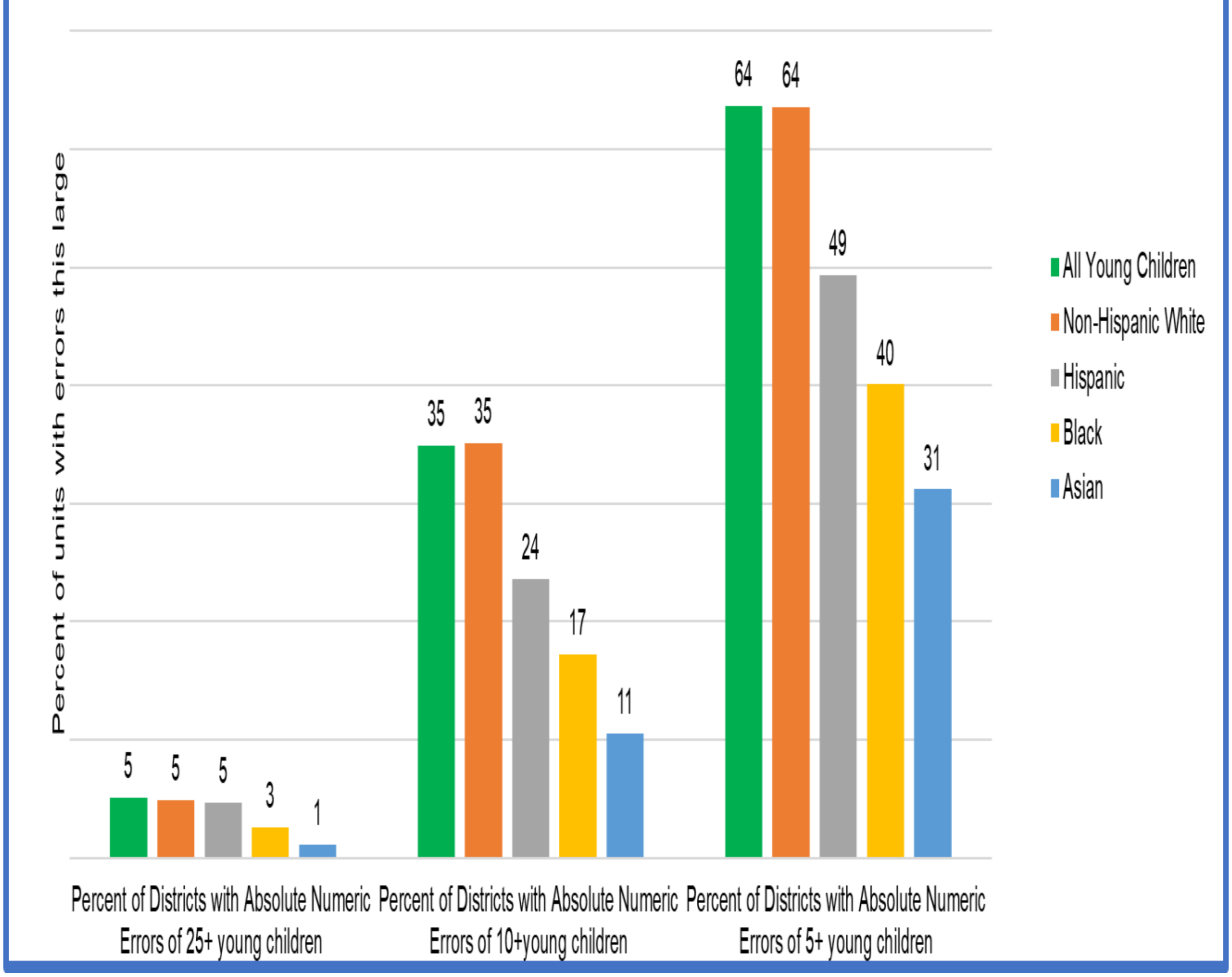


I use three benchmarks for large absolute *numeric* errors. The 5 persons and 10 persons categories of error have been used in other publications. I added the 25 persons plus category to look at the most extreme errors. Errors of 25 or more young children are likely to be very problematic in many Unified School Districts.

Figure 2 shows 64 percent of the Unified School Districts had errors of 5 young children or more for young children of all races but the figures for minority groups are smaller: 49 percent for Hispanic young children, 40 percent for Black young children, and 31 percent for Asian young children.

In Figure 2, in each category of absolute *numeric* errors (5 young children, 10 young children, and 25 young children), there are many fewer districts that have this level of error for Hispanic, Black, and Asian young children than there are districts that have this level of error for all young children or Non-Hispanic White young children. This is because these are generally smaller populations.

Figure 2. Distribution of Absolute *Numeric* Errors for Population Ages 0 to 4 for Unified School Districts by Race and Hispanic Origin



There are relatively few Unified School Districts with very large absolute *numeric* errors. Only 5 percent of Unified School Districts have errors of 25 young children or more, compared to 5 percent of Hispanic young children, 3 percent for Black young children, and 1 percent for Asian young children. For those districts that have errors of 25 percent or more because of the application of DP, the results are likely to be a substantial problem

The national numbers shown above mask a lot of variation across states. Table 4 shows states ranked on two key measures of accuracy (mean absolute *numeric* error and mean absolute *percent* error) for Unified School Districts. The mean absolute *percent* error ranges from a low of 3.4 for Vermont to a high of 16.9 percent in California. The mean absolute *percent* error for states ranges from a low of 0 for Hawaii (Hawaii only has one unified school district) to a high of 15.8 for Montana.

Table 4. States Ranked on Mean Absolute Numeric Error and Mean Absolute Percent Error for Children ages 0 to 4 by Unified School Districts

Rank*		Average of Absolute *Numeric* Difference		Rank*		Average of Absolute *Percent* Difference
1	California	16.9		1	Montana	15.8
2	Delaware	13.9		2	Alaska	14.0
3	Arizona	13.3		3	Maine	11.2
4	Hawaii	13.0		4	Vermont	10.1
5	Michigan	11.5		5	Washington	8.1
6	New York	11.2		6	North Dakota	7.4
7	Florida	10.1		7	Nebraska	6.6
8	Texas	10.1		8	South Dakota	6.6
9	Mississippi	9.9		9	Oregon	6.2
10	Arkansas	9.8		10	Colorado	6.1
11	Washington	9.8		11	New Mexico	5.8
12	Illinois	9.8		12	Oklahoma	5.3
13	Oregon	9.7		13	Idaho	5.2
14	South Carolina	9.6		14	Texas	5.1
15	Missouri	9.5		15	Kansas	4.9
16	Oklahoma	9.4		16	Indiana	4.6
17	North Carolina	9.3		17	Wyoming	4.4
18	Utah	9.3		18	Iowa	4.2
19	Minnesota	9.3		19	New Hampshire	4.1
20	Wisconsin	9.2		20	Missouri	4.1
21	Ohio	9.0		21	Minnesota	3.4
22	Idaho	8.6		22	Ohio	3.0
23	Iowa	8.6		23	Wisconsin	3.0
24	New Mexico	8.5		24	New York	2.9
25	Louisiana	8.3		25	Arkansas	2.9
26	Colorado	8.2		26	Illinois	2.7
27	Indiana	7.9		27	Michigan	2.5
28	Kansas	7.9		28	Arizona	2.1
29	Maryland	7.9		29	California	1.8
30	Tennessee	7.8		30	New Jersey	1.6
31	Georgia	7.8		31	Nevada	1.5
32	Wyoming	7.8		32	Mississippi	1.4
33	Alabama	7.7		33	Kentucky	1.3
34	Pennsylvania	7.5		34	Pennsylvania	1.0
35	Connecticut	7.4		35	Utah	0.9
36	Nebraska	7.3		36	Massachusetts	0.9
37	Virginia	7.2		37	Tennessee	0.8
38	Nevada	7.2		38	Virginia	0.8
39	South Dakota	7.0		39	Alabama	0.7
40	Kentucky	7.0		40	South Carolina	0.7
41	Massachusetts	6.5		41	Georgia	0.7
42	New Jersey	6.2		42	Rhode Island	0.7
43	Montana	5.6		43	Connecticut	0.6
44	North Dakota	5.3		44	West Virginia	0.6
45	New Hampshire	5.1		45	Delaware	0.6
46	West Virginia	5.0		46	North Carolina	0.5
47	Alaska	4.9		47	Louisiana	0.4
48	Rhode Island	4.8		48	Florida	0.3
49	Maine	3.6		49	Maryland	0.2
50	Vermont	3.4		50	Hawaii	0.0
U.S. Average		9				3.7

Source: Authors analysis of Demonstration Product released by the Census Bureau August 25, 2022 after process by IPUMS NIHGIS at the University of Minnesota

* Ranking is based on unrounded data.

Analysis for Age 4

In the Demonstration Product released in August 2022, the Census Bureau provided data by single year of age and sex for the population under age 20. I analyze this data for age 4 for Unified School Districts. I selected age 4 because that is often used by school systems to predict the number of kindergarteners to expect in the following school year. I do not see any reason why the metrics for age 4 would be much different than the metrics for any other single year of age.

Table 5 provides the key metrics for the comparison of age 4 in Unified School Districts in the 2010 Census file with and without DP. Districts with no people age 4 in the DP or SF file were not used in the analysis. The mean absolute *numeric* error was 9 and the mean absolute *percen*t error was 11 percent for age 4

A large share of Unified School Districts had large errors in both *numeric* and *percent* terms. About three out of five (59 percent) of Unified School System had absolute *numeric* errors of 5 or more children and 52 percent of Unified School Districts had absolute *percent* errors of 5 percent or more for children age 4.

With errors of this magnitude for single year of age, one has to wonder if this data is worth producing. This is particularly true for smaller districts where the errors are likely to be larger percentage-wise. It is not clear how users are supposed to manage data with this degree of uncertainty

.

Table 5. Unified School District Error* Metrics for Age 4

Number of Units in Analysis	10,782
Mean number of children age 4 in Summary File	376
Mean Absolute *Numeric* Error	9
Mean Absolute *Percent* Error	11
Percent of units with Absolute *Numeric* Error 5+ children age 4	59
Percent of units with Absolute *Percent* Error 5%+ **	52

Source: Author's analysis of Demonstration Product released by the Census Bureau on August 25, 2022 after processing by IPUMS NHGIS at the, University of Minnesota www.nhgis.org

* In this paper, errors reflect the difference between the 2010 Census data without and with DP injected.

Data in this table does not include Puerto Rico or geographic units with zero population age 4 in 2010 Summery File or DP-Infused file.

** analysis is based on figures rounded to two decimal points

Data for Places

Census Places are geographic units used by the U.S. Census Bureau to publish data. They range from Places with millions of people such as Los Angeles and New York City, to the smallest villages and towns.

Places include both incorporated Places and Census Designated Places (CDPs). There are a little more than 29,000 Places for which the infusion of DP data was produced in the August 16, 2022 (DP Demonstration Product) and most of them (over 19,000) are Incorporated Places rather than Census Designated Places (CDPs). Incorporated Places are legally bounded entities such as cities, boroughs, towns, or villages (names may vary depending on the state). Census Designated Places (CDPs) are statistical entities used in the Census. They are unincorporated communities where there is a concentration of population, housing, and commercial structures and they are identifiable by name. There are nearly 10,000 CDPs for 2010 Census data.

Cities, villages, and towns might want to know about the number of young children in their area for things like planning youth activities, child facilities, and day care centers. The preschool-age population is also useful for forecasting future school enrollments.

Table 1 shows the mean absolute *numeric* error for Places was 5 and the mean absolute *percent* error was 13 percent. The high percent error is not surprising because many of these Places are small. There were 1,422 Places where the number of young children was less than 100, and 9,012 Places where the number of young children was less than 500, based on the 2010 Summary File.

Figure 3 shows the distribution of Places by absolute *percent* error using the same thresholds used for Unified School Districts. The data in Figure 3 shows that almost half (46 percent) of Places had absolute *percent* errors of 5 percent or more for the young child population and 15 percent had absolute *percent* errors of 25 percent or more. Since Places are generally smaller (in population size) than Unified School Districts, it is not surprising that the percentages are larger for Places than for Unified School Districts.

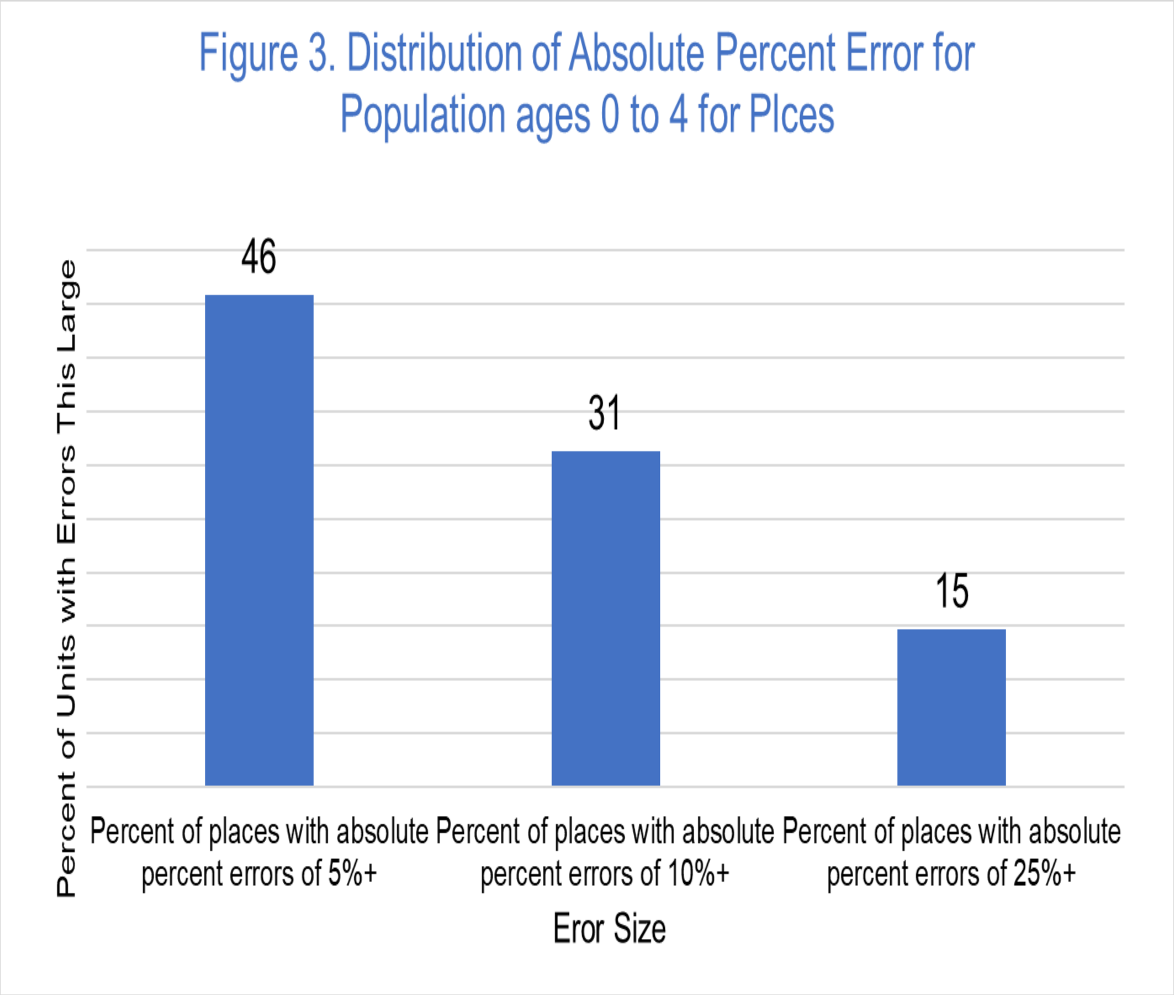


Figure 4 shows the distribution of Places by absolute *numeric* errors using the same categories as Figure 2. Data show 38 percent of the Places had absolute *numeric* errors of 5 or more young children, and only 2 percent had absolute *percent* errors of 25 or more young children.

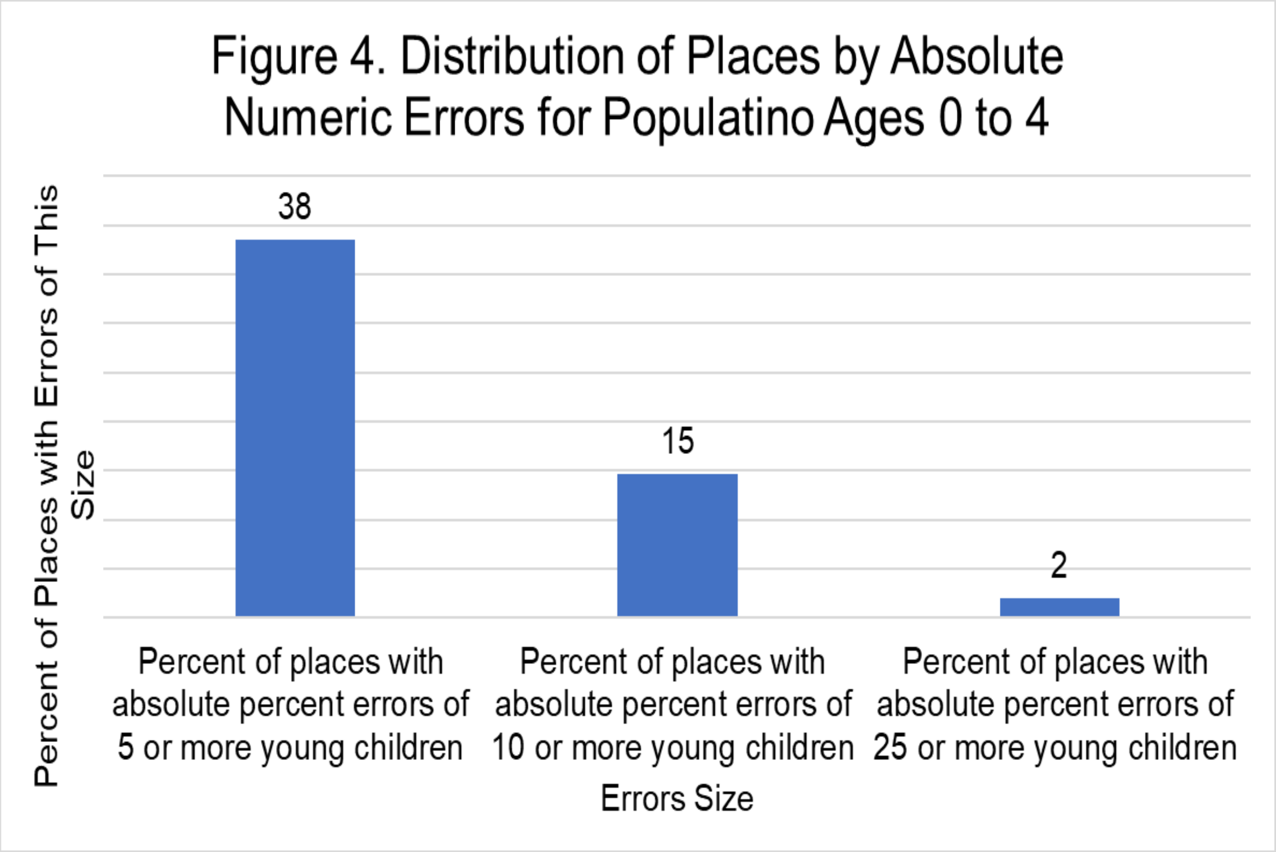


Table 6 shows states ranked on the percent of places in a state with absolute *percent* errors of 5 percent or more. Data for errors of 10 percent or more and 25 percent or more are also provided in the Table 6.

There is a lot of variation across the states. For example, 67 percent of the Places in Nebraska had absolute *percent* errors of 5 percent or more, compared to 25 percent of Places in New Jersey.

Table. 6 States Ranked by Percent of Places in State with Absolute Percent Errors of 5 Percent or More for Ages 0 to 4

Rank*	State	Number of Places in State	Errors of 5% +	Errors of 10% +	Errors of 25% +
1	Nebraska	543	67	50	26
2	New Mexico	419	64	50	28
3	Vermont	116	63	49	27
4	North Dakota	346	63	51	30
5	South Dakota	353	62	47	26
6	Montana	341	62	51	34
7	Wyoming	184	60	48	28
8	Alaska	310	60	45	28
9	West Virginia	397	60	42	20
10	Oklahoma	708	57	42	20
11	Arizona	429	56	41	23
12	Maine	130	55	35	5
13	New Hampshire	96	55	38	22
14	Kansas	648	55	41	20
15	Iowa	976	54	36	18
16	Missouri	994	51	38	17
17	Arkansas	529	51	33	14
18	Pennsylvania	1742	50	34	15
19	Colorado	435	48	35	20
20	Nevada	123	48	36	22
21	Minnesota	895	46	31	14
22	North Carolina	733	46	30	12
23	Massachusetts	242	46	29	8
24	Virginia	590	45	34	16
25	Kentucky	520	45	27	13
26	Idaho	215	45	27	10
27	Wisconsin	761	44	28	12
28	South Carolina	392	44	27	12
29	New York	1181	43	27	10
30	Alabama	574	43	30	13
31	Texas	1714	43	29	14
32	Oregon	365	43	32	19
33	Delaware	75	43	31	15
34	Utah	318	42	28	10
35	Washington	610	42	30	15
36	Indiana	677	42	25	11
37	Ohio	1198	40	26	11
38	Illinois	1360	40	26	10
39	Michigan	684	39	26	10
40	Maryland	507	38	32	18
41	Rhode Island	34	38	35	18
42	Louisiana	472	38	22	8
43	California	1466	38	26	15
44	Georgia	622	38	23	9
45	Tennessee	427	38	22	8
46	Mississippi	360	36	20	8
47	Hawaii	150	35	21	8
48	Connecticut	142	35	20	4
49	Florida	908	31	19	8
50	New Jersey	536	25	17	9
	U.S. Total	28547	46	31	15

Source: Author's analysis of Demonstration Product data released by the Census Bureau on August 25, 2022 after being processed by IPUMS NHGIS at the University of Minnesota www.nhgis.org

Data in this table does not include Puerto Rico or geographic units with zero population age 0 to 4 in 2010 Summary File or DP-infused file.

* in this paper errors reflect the difference between the 2010 Census data without and with DP injected.

** The Census Bureau calls this measure Mean Absolute Error. I include the word "Numeric" to distinguish it from Mean Absolute Percent Error.

Table 7 shows states ranked on the percent of places in the state with absolute *numeric* errors of 5 or more young children. Data for 10 or more and 25 young children or more are also shown in the table. There is a lot of variation among the states. For example, 69 percent of places in Massachusetts have absoltue *numeric* errors of 5 or more young children compared to12 percent of North Dakota.

Table 7. States Ranked by Percent of Places in State with Absolute Numeric Errors

Rank*	State	Number of Places in State	Errors of 5+ young children	Errors of 10+ young children	Errors of 25+ young children
1	Massachusetts	242	69	43	7
2	Hawaii	150	65	37	9
3	Maine	130	65	29	6
4	Connecticut	142	64	35	9
5	California	1,465	58	28	4
6	New Hampshire	96	58	26	4
7	Florida	908	55	30	5
8	Rhode Island	34	53	18	0
9	Virginia	590	51	25	7
10	Maryland	507	49	22	6
11	Vermont	116	49	20	1
12	Washington	610	49	23	3
13	New Jersey	536	49	22	3
14	Arizona	429	48	23	3
15	New York	1,181	48	23	3
16	Utah	318	46	17	2
17	Nevada	123	46	20	4
18	Texas	1,714	44	14	2
19	South Carolina	392	43	19	3
20	Michigan	684	43	14	2
21	Delaware	75	43	19	8
22	North Carolina	733	42	15	1
23	Louisiana	472	42	14	3
24	New Mexico	419	40	16	1
25	Pennsylvania	1,742	40	18	2
26	Oregon	365	39	14	2
27	Tennessee	427	38	12	1
28	Georgia	622	38	14	1
29	Colorado	435	37	14	2
30	Ohio	1,198	36	11	1
31	Wisconsin	761	34	9	1
32	Alabama	574	34	11	1
33	West Virginia	397	34	14	1
34	Mississippi	360	33	11	1
35	Kentucky	520	33	9	1
36	Montana	341	32	12	1
37	Illinois	1,360	31	9	1
38	Oklahoma	708	31	7	0
39	Indiana	677	31	9	1
40	Alaska	310	27	9	2
41	Wyoming	184	27	7	0
42	Idaho	215	26	6	1
43	Minnesota	895	25	5	0
44	Missouri	994	25	6	0
45	Arkansas	529	22	4	0
46	South Dakota	353	20	6	0
47	Nebraska	543	20	3	0
48	Kansas	648	19	4	0
49	Iowa	976	18	2	0
50	North Dakota	346	12	2	0
	U.S. Total	28,546	38	14	2

Source: Author's analysis of Demonstration Product data released by the Census Bureau on August 25, 2022 after being processed by IPUMS NHGIS at the University of Minnesota www.nhgis.org

Data in this table does not include Puerto Rico or geographic units with zero population age 0 to 4 in 2010 Summary File or DP-infused file.

* in this paper errors reflect the difference between the 2010 Census data without and with DP injected.

DC is not included in the state data but is included in the county data

Discussion

It is clear that the introduction of DP into the 2020 Census has caused a lot of controversy. I have been following the U.S. Census since 1970, and I do not remember any issue that has caused as much discussion, concern, and debate among data users as the decision to implement DP in the 2020 Census.

Below I review a couple of issues regarding DP that were not addressed in my analysis but may impact stakeholder's view of DP

Block-Level Data

Blocks are the smallest geographic unit used in the Census and there are about 8 million blocks in the 2020 Census but only about 6 million are occupied. The average block has a total population of about 41 people and about 3 young children. The small population size of blocks makes them susceptible to large percent errors when random numbers are injected with DP.

Assessment of Census accuracy using the two standard Census Bureau methods (Demographic Analysis and Post-Enumeration Survey) are not available at the sub-state level. But the DP Demonstration Product allows one to look at errors attributable to DP for all levels of Census geography down to the census block level and there are some very troublesome issues regarding the use of DP at the census block level.

There are two broad perspectives on the error DP injects into census blocks. One perspective is that data for census blocks are among the most important data supplied by the Decennial Census, and they need to be as accurate as possible. One of the primary purposes of the Decennial Census is to provide comparable population figures for small areas across the country. To the best of my knowledge, there is no other data source that provides demographic data for all the blocks in the country other than the Decennial Census. Consequently, census accuracy for blocks is especially important. O'Hara (2022) makes a strong case for why block level data are important in terms of creating special or custom districts. The need for such data is often not apparent until well after the Census data has been collected and reported.

Another perspective holds that blocks are typically aggregated into larger units like congressional districts, cities, and counties and in those aggregations the random error injected into individual blocks cancel each other out and produce relatively accurate data for larger units. From this perspective, errors at the block level are not so important.

Regarding the usability of block level data, the Census Bureau (Devine 2022, slide 17) recently stated, "Block-level data are fit-for-use when aggregated into geographically contiguous larger entities. They are not intended to be fit-for-use as a unit of analysis."

I do not think there is any dispute that the error injected by DP for blocks produces a relatively high absolute *percent* error and that these errors typically cancel each other out when blocks are aggregated into larger areas. Because the error is random, the amount of error does not become cumulative. It is an open question about how important census block level data are for making decisions.

One problem with use of DP for small areas is the implausible or impossible results produced. For example, more than 163,000 blocks have children (population age 0 to 17) but no adults (population age 18 and over) after DP is applied compared to just 82 such blocks before DP was applied (U.S. Census Bureau 2022d). Many such cases are highly unlikely and raise questions about who these children are living with if there are no adults in their household. The Census Bureau (2022d) offers several other examples of implausible or impossible results in the data after DP is applied.

It is not clear to me exactly what statistical problems might be caused by these results, but they undermine the veracity of the census data broadly. A high number of improbable results is identified as a problem of “legitimacy” rather than statistical accuracy by Hogan (2021) and is likely to undermine the confidence the public has in the Census results. When data users see highly implausible results like the large number of blocks with children and no adults, they often wonder what other errors are in the data that are not so apparent.

Despite the statement by the Census Bureau about using block-level data and misgivings among some demographers about the quality of census block data, many data users routinely use the block level data, either because they do not realize the level of potential errors, or because it is the best (or only) data they have at that level of geography.

The data indicate the average percent errors for census blocks is relatively high but does not address how often block-level data are used in decision-making. Readers may have their own answer to that question.

Breaking the Link Between Child and Parents

The production of many blocks where there are children, but no adults may be related to the link between children and adults in a household that is broken when 2020 Disclosure Avoidance System(DAS) with Differential Privacy (DP) was applied to the DHC file. DP is administered to children (population age 0 to 17) and parents (population age 18 and over) independently, so it may eliminate the adults in a household that has children by randomly subtracting data from the number of adults. If the processing retained the link between young children and their parents in a household, it is doubtful that there would be such a high number of blocks with children and no adults.

This statistical disconnection of children and parents is an on-going concern and is likely to have important impacts in later Census products which have more detailed data on young children.[3] For example the connection between children and parents is critical for a lot of data from the American Community Survey. Child poverty is probably the single most important measure of child well-being and determining poverty status requires linking a child to the income of the adults in the households.

The Census Bureau says it will use a different method of DP in the Detailed Demographic and Housing File which will retain the connection between children and parents. Hopefully, that will alleviate concerns. But data that links children and adults in the Detailed Demographic and Housing file will not be available until late 2023 or 2024. That is getting very close to the date (2025) the Census Bureau said it might start applying DP to the American Community Survey (ACS). Translating the application of

[3] It is my understanding that the use of DP does not necessarily require the disconnect between children and parents in a household. The break between children and parents in the redistricting file and the DHC is a result of the particular DP-related processing chosen by the Census Bureau.

DP from the Census to the ACS, is likely to be a difficult process because the ACS is a sample survey rather than a census and the ACS measures more than 40 topics.

Accuracy and Equity

The focus of this report is on census accuracy, but the differential accuracy revealed in my analysis raises the issue of equity. Equity in terms of data provision has become a more visible aspect of data collection and reporting in the federal government recently (White House Equitable Data working Group 2022). According to the U.S. Census Bureau (2021e, pages 1) " The Census Bureau has an ongoing commitment to producing data that depict an accurate portrait of America, including its underserved communities." Data equity has become a part of broader equity questions. This suggests all results should be examined through the lens of equitable data.

In terms of equity, Table 3 shows substantial differential accuracy for Unified School Districts by race and ethnicity after DP is applied in terms of absolute *percent* errors. For Hispanic young children, the mean absolute *percent* error was 28, for Black young children the mean absolute *percent* error was 35, and for Asian young children was 45, compared to 5 for non-Hispanic white children. What does this say about the equity of using the DP method?

There is already differential accuracy in census results before DP is applied but it may be the case that DP exacerbates such inequities. Is it fair to inject as much error for groups that already have a lot of error in census data as for those groups that do not have much error? Did the Census Bureau examine equity concerns when they decided to use DP in the 2020 Census?

Selection of a Disclosure Avoidance System and Public Trust

Disclosure avoidance is not just a statistical issue and examining it only from a statistical perspective may be problematic. Another dimension for assessing alternative DAS methods is the extent to which a given DAS method undermines public trust in the Census data and the Census Bureau itself. There has been a great deal of concern about the erosion of public trust in the Census Bureau recently. According to the National Academy of Sciences, Engineering and Medicine panel assessing the 2020 Census (2022, page 6),

"We are very concerned, based on presentations to the panel and our knowledge of reactions to previous demonstration data, that the Census Bureau's adoption of differential privacy-based disclosure avoidance has increased the level of public mistrust in the 2020 Census and the Census Bureau itself."

A recent statement from the Federal States Cooperative Program for Population Estimate (FSCPE 2022) states, "Differentially private algorithms have appropriate applications, but they are not a panacea. The evidence and experience to date indicate that they are not capable of handling the complexity of the nation's political and statistical geography and hence do not provide usable data for key constituents."

In their review of the impact DP has had on the Census Bureau credibility and trust among data users, Boyd and Sarathy (2022, page 1) state, "We argue that rebuilding trust will require more than technical repairs or improved communication: it will require reconstructing what we identify as a “statistical imaginary.”

Summary

This report provides information on the accuracy of DP-infused data and provides a profile of the likely errors for young children that will be seen in data for in the 2020 Census if the Census Bureau uses the privacy protection parameters reflected in the August 2022 Demonstration Product.

It is important to note that the analysis provided in this paper is just a sample of analyses that could be done. But I believe the data analyzed in this study a relatively good sample of the broader implications of using a DAS method with DP in the Demographic and Housing Characteristics file with the privacy protection parameters used in this Demonstration Product.

The question that is not addressed in the previous sections is whether the level of error reflected in this analysis would make 2020 Census for data on young children “unfit for use.” Each person will probably have a different answer to how much error in census data for young children is too much error.

Like all disclosure avoidance systems, the use of DP involves a trade-off between privacy protection and census accuracy. There have always been errors in the Census data, but in the 2020 Census, the Census Bureau is trying to decide how much additional error to add to the data in order to enhance privacy protection. By setting privacy parameters, the Census Bureau has control over the level of accuracy and level of privacy protection in the 2020 Census.

Given this balancing act, it would be useful to have more information about metrics on privacy protection. It would be helpful if we could compare the metrics of accuracy like those in this report to metrics of privacy protection in the August 2022 Demonstration Product. I see many measures of accuracy based on the Demonstration Product. However, I do not see any privacy protection metrics produced by the Census Bureau nor do I see a way to explore the privacy protection aspect with the Demonstration Product. It seems the balance of accuracy and privacy protection is the key reason for using a given disclosure avoidance system but without metrics for privacy protection I am not sure how to do that. When I have asked experts about the level of privacy protection afforded by an Epsilon of 19.6 in the redistricting data in terms I can understand it seems like I always get a variation of “it depends.” But no metrics.

On the other hand, the problems that are likely to be caused by inaccurate census data on young children are clearer to me. The data in this paper, and many other analyses, provide a rich set of metrics showing the magnitude of error DP injects into Census data and I can envision problems such errors might cause.

When the number of young children in a school district is under-reported by 5 or 10 percent, that could have big implications for their funding and when the number of young children in a community is off by 10 percent or more, that could impact planning in ways that waste taxpayer money and undermine quality education for young children. If the number of young children reported in the Census for a Unified School District is 10 percent too low, it may not automatically translate into 10 percent less money for that jurisdiction. But there is a strong link between underreporting the number of young children and the loss of money in a general sense.

In addition to the money distributed on the basis of census-derived data, Census data are used for many decisions in the public and private sector. The more errors there are in the data and the larger the errors in the data, the less likely those decisions will be correct ones.

Given the level of errors in Unified School Districts and Places using the privacy protection level in the most recent DP Demonstration Product, and the lack of clear evidence or measurements about the level or impact of privacy loss, I recommend that the Census Bureau increase the level of accuracy used in the DHC to provide more accurate small area data for young children. And reduce or eliminate large errors caused by application of DP.

References

Bouk, D. and Boyd, D. (2021*). Democracy’s Data Infrastructure.; The technologies of the U.S. Census*. https://knightcolumbia.org/content/democracys-data-infrastructure

Boyd. D. (2019). “Balancing Data Utility and Confidentiality on the 2020 US_Census,” Data and Society, https://datasociety.net/library/balancing-data-utility-and-confidentiality-in-the-2020-us-census/ .

Boyd, D. and Sarathy, J. (2022) “Differential Perspectives: Epistemic Disconnects Surrounding the US Census Burau’s Use of Differential Privacy,” *Harvard Data Science Review* (forthcoming)

Committee on National Statistics (2019). “Workshop on 2020 Census Data Products: Data Needs and Privacy Considerations,” presentations are available at https://www.nationalacademies.org/event/12-11-2019/workshop-on-2020-census-data-products-data-needs-and-privacy-considerations .

Cropper, M., McKibben, J, and Stojakovic, Z. (2021). The Importance of Small Area Census Data for School Demographics, Count all Kid website https://ednote.ecs.org/counting-all-kids-how-the-census-impacts-education/

Federal State Cooperative Population Estimates (FSCPE) (2022). “Letter to Census Bureau Director Robert Santos,” https://docs.google.com/forms/d/e/1FAIpQLScU7bK9yIAy9YV-WIVjIJhx-b05-IB2el8M47Cg1jZu3Sa5tA/viewform

Hogan, H. (2021). “The History of Assessing Census Quality, Presentation at 2021 Population of Association of America Conference, May 5, 2021.

Hogan, H. (2021). “The History of Assessing Census Quality, Presentation at 2021 Population of Association of America Conference, May 5, 2021.

Hotz, J. and Salvo J. (2020). Addressing the Use of Differential Privacy for the 2020 Census: Summary of What We Learned from the CNSTAT Workshop. https://www.apdu.org/2020/02/28/apdu-member-post-assessing-the-use-of-differential-privacy-for-the-2020-census-summary-of-what-we-learned-from-the-cnstat-workshop/ ,

McElrath, K. Bauman, K., and Schmidt, E (2022) Preschool Enrollment in the United states,: 2055 to 2019,” U.S. Census Bureau https://www.census.gov/content/dam/Census/newsroom/press-kits/2021/paa/paa-2021-presentation-preschool-enrollment-in-the-united-states.pdf

Nagle, N. and Kuhn, T. (2019). “Implications for School Enrollment Statistics.” https://www.nationalacademies.org/event/12-11-2019/workshop-on-2020-census-data-products-data-needs-and-privacy-considerations.

National Academy of Sciences, Engineering and Medicine, (2022). *Understanding the Quality of the 2020 Census, Interim Report* , Washington Dc. The National Academy Press, https://nap.nationalacademies.org/catalog/26529/understanding-the-quality-of-the-2020-census-interim-report

O’Hara, A. (2022) presentation at Analysis of Census Noise Measurements Workshop, April 28-29, Rutgers University.

O’Hare, W.P. (2019). “Assessing 2010 Census Data with Differential Privacy for Young Children,”” https://www.nationalacademies.org/event/12-11-2019/workshop-on-2020-census-data-products-data-needs-and-privacy-considerations .

O’Hare W. P. (2020a). “Many States Use Decennial Census Data to Distribute State Money,” The Census Project Website https://thecensusproject.org/2020/01/09/many-states-use-decennial-census-data-to-distribute-state-money/

O’Hare, W.P (2020b). “Implications of Differential Privacy for Reported Data on Young children in the 2020 U.S. Census,” Posted on Count All KIDS Website Implications-of-Differential-Privacy-for-kids-11-17-2020-FINAL-00000003.pdf (myftpupload.com) .

O’Hare, W.P. (2021). “Analysis of Census Bureau’s August 2021 Differential Privacy Demonstration Product: Implications for Data on Children,” Count All Kids website November *https://countallkids.org/resources/analysis-of-census-bureaus-august-2021-differential-privacy-demonstration-product-implications-for-data-on-children/*

O’Hare, W. P. (2022a). “New Census Bureau Data Show Young Children Have a High Net Undercount in the 2020 Census, “ Posted on Count All Kids website , March, https://countallkids.org/resources/new-census-bureau-data-show-young-children-have-a-high-net-undercount-in-the-2020-census/

O’Hare , W. P. (2022b). “U.se of the American Community Survey Data by State Child Advocacy Organizations.” Count All Kids website, https://countallkids.org/resources/use-of-the-american-community-survey-data-by-state-child-advocacy-organizations/

O’Hare W. P and A. Rashid (2022). Selected Federal Programs that Use Figures for the Population Ages 0 to 5 for Distribution of Federal Funds to States and Localities, Posted on Count All Kids website July 5 https://countallkids.org/selected-federal-programs-that-use-the-population-size-for-ages-0-to-5-for-the-distribution-of-federal-funds-to-states-and-localities/

O'Hare, W.P. (2022c). "Analysis of Census Bureau's March 2022 Differential Privacy Demonstration Product: Implications for Data on Young Children," Posted on the Count All Kids website, https://countallkids.org/resources/analysis-of-census-bureaus-march-2022-differential-privacy-demonstration-product-implications-for-data-on-young-children/

Reamer, A. (2020). Counting for Dollars, George Washington University https://gwipp.gwu.edu/counting-dollars-2020-role-decennial-census-geographic-distribution-federal-funds .+

The Annie E. Casey Foundation (2018). *KID COUNT DATA BOOK 2018*, https://www.aecf.org/resources/2018-kids-count-data-book

U.S. Census Bureau (2018), "Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing," THE RESEARCH AND METHODOLOGY DIRECTORATE, Mc Kenna, L. U.S. Census Bureau, Washington DC., https://www.census.gov/content/dam/Census/library/working-papers/2018/adrm/Disclosure%20Avoidance%20for%20the%201970-2010%20Censuses.pdf .

U.S. Census Bureau (2019). "2010 Demonstration Data Products," U.S. Census Bureau, Washington DC., October, https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2010-demonstration-data-products.html .

U.S. Census Bureau (2020a). 2020 Census Disclosure Avoidance Improvement Metrics, U.S Census Bureau, Washington DC., August18, https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/disclosure-avoidance-system/2020-03-18-2020-census-da-improvement-metrics.pdf?# .

U.S. Census Bureau (2020b). "2020 Census Data Products and the Disclosure Avoidance System", Hawes M. and Garfinkel. S. L., Planned presentation at the Census Scientific Advisory Committee meeting, August26.

U.S. Census Bureau (2020c). "DAS Updates, U.S Census Bureau," Hawes M. June 1 Washington DC., https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/disclosure-avoidance-system/2020-06-01-das-updates.pdf?# .

U.S. Census Bureau (2020d). "Disclosure Avoidance and the Census," Select Topics in International Censuses, U.S. Census Bureau, October 2020. https://www.census.gov/library/working-papers/2020/demo/disclos-avoid-census.html .

U.S. Census Bureau (2020e). “Disclosure Avoidance and the 2020 Census, U.S. Census Bureau,” Washington DC., Accessed November 2, https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html .

U.S. Census Bureau (2020f). “Error Discovered in PPM,” U.S. Census Bureau, Washington DC. https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-updates.html .

U.S. Census Bureau (2020g). “2020 Disclosure Avoidance System Updates,” U.S. Census Bureau, Washington DC., https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-updates.html .

U.S. Census Bureau (2021a). School Enrollment in the United States: October 2019 - PowerPoint Presentation (census.gov Detailed Tables, School Enrollment in the United States: October 2019 - Detailed Table 1, FEBRUARY 02, 2021.

U.S. Census Bureau (2021b). Developing the DAS: Demonstration Data and Progress Metric, https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/2020-das-development.html .

U.S. Census Bureau (2021c). “Differential Privacy 101.” Webinar May 4, 2021, Michael Hawes. https://www.census.gov/data/academy/webinars/2021/disclosure-avoidance-series/differential-privacy-101.html

U.S. Census Bureau (2021d). “Disclosure Avoidance for the 2020 Census: An Introduction,” U.S. Census Bureau, Washington, DC. November https://www.census.gov/library/publications/2021/decennial/2020-census-disclosure-avoidance-handbook.html

U.S. Census Bureau (2021e). “Advancing Equity with Census Bureau Data.” Census Bureau Blog, November 2, 2021, Ron Jarmin , Acting Director Advancing Equity with Census Bureau Data

U.S. Census Bureau (2021f). “Disclosure Avoidance for the 2020 Census: An Introduction,” November 2021, U.S. Census Bureau, Washington DC https://www.census.gov/library/publications/2021/decennial/2020-census-disclosure-avoidance-handbook.html

U.S. Census Bureau (2022a). “Understanding Disclosure Avoidance- Related Variability in the 2020 Census Redistricting data, “ U.S. Census Bureau, Washington DC. January 28. https://www.census.gov/library/fact-sheets/2022/variability.html

U.S. Census Bureau (2022b). “Revised Data Metrics for 2020 Disclosure Avoidance,” U.S. Census Bureau, Washington DC.

U.S. Census Bureau (2022c) Post-Enumeration Survey and Demographic Analysis Help Evaluate 2020 Census Results, August10 , Census Bureau Releases Estimates of Undercount and Overcount in the 2020 Census

U.S. Census Bureau (2022d). Detailed Summary Metrics , https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/02-Demographic_and_Housing_Characteristics/2022-08-25_Summary_File/2022-08-25_Detailed_Summary_Metrics_Overview.pdf

U.S. Census Bureau (2022e) "Just Released: New Demonstration Data for the DHC; webinar August 31,

U.S. Census Bureau (2022f). Summary of Feedback on DHC Demonstration Data JUNE 23, 2022 https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/newsletters/summary-of-feedback-on-dhc-demonstration-data.htm

U.S. Census Burau (2022g) Just Released: New Demonstration Data for the DHC; Webinar August 31, August 25, https://www.census.gov/programs-surveys/decennial-census/decade/2020/planning-management/process/disclosure-avoidance/newsletters/New-2010-DHC-Demonstration-Data-Coming-August-25-Webinar-August-31.html

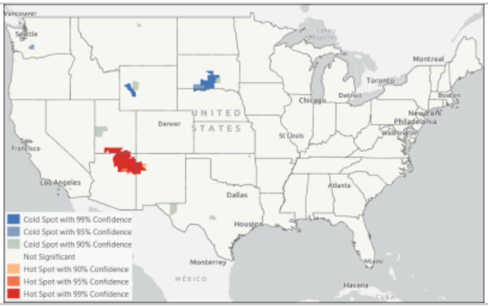
U.S. General Accountability Office (2020). "COVID-19 Presents Delays and Risks to Census Counts," U.S. General Accountability Office, Washington, DC., https://www.gao.gov/products/GAO-20-551R .

Vink, J. (2019). "Elementary School Enrollment," https://www.nationalacademies.org/event/12-11-2019/workshop-on-2020-census-data-products-data-needs-and-privacy-considerations .

White House Equitable Data Working Group (2022) "A Vision for Equitable Data : Recommendations from the Equitable Data Working Group," https://www.whitehouse.gov/wp-content/uploads/2022/04/eo13985-vision-for-equitable-data.pdf

4. Lauren Scott Griffin, Spatial Analytics, ESRI

In response to your call for feedback, I have prepared this spatial analysis of the v20220825 demonstration data: https://arcg.is/15DvCH3



Census Demonstration Data: v20220825

Spatial analysis

arcg.is

I've identified several issues I hope you will address:

The range in the injected noise for different population groups isn't very similar, meaning the data distortion due to DAS will disadvantage some population groups more than others.
For some population groups the relationship between the DAS Offsets (DP-SF1) and Total Population (or total population for a particular population group), is not zero (R ≠ 0) and/or there is evidence of heteroskedasticity. This means tracts with either larger or smaller populations may be disadvantaged due to more injected noise (more data distortion).

While overall the population overestimates and underestimates balance out when viewed in a spreadsheet, there is definite evidence of spatial clustering when viewed and analyzed on a map. This is concerning, and identified as a problem for all the population groups I tested (all except Total Population). For the AIAN population, for example, there is statistically significant clustering of underestimates in South Dakota, places where Native Americans may already be resource challenged. Spatial clustering of population underestimates would result in underfunding or underrepresentation for programs based on population counts.

The impact of large over or underestimates is especially problematic where population totals are small. Some of the absolute errors are as high as 2245 percent. This seems unacceptable.

Thanks for considering my feedback as you move forward.

Lauren

Analysis of Unified School Districts and Places with Large Errors for the Population Ages 0 to 4 Caused by Application of Differential Privacy

By

Dr. William P. O’Hare

In my analysis of the DP Demonstration Product released by the Census Bureau on August 25, 2022 (O’Hare 2022) I found many Unified School Districts and Places[1] where DP introduced large errors for the population ages 0 to 4. The population ages 0 to 4 is referred to as young children in this paper. Large errors are defined in this paper as absolute *percent* errors of more than 25 percent or absolute *numeric* errors of more the 25 young children.[2]

This paper focuses on those units that have an absolute *percent* error of 25 percent or more and those units that have absolute *numeric* errors of 25 or more young children (age 0 to 4). In a recent presentation, the Census Bureau (2022, slide 3) describes DP as, “DP inserts small differences into counts of people and households, making it very difficult to identify people.” In my opinion, errors of 25 percent or more are not “small”.

Numeric errors reflect the difference in 2020 Census results with and without DP. Absolute *percent* difference converts the *numeric* difference to a relative difference by dividing by the 2010 Census Summary File value and multiplying by 100.. It may be a

[1] These include incorporated Places as well and Census Designated Places.

[2] Errors in this application are the difference in the 2010 Census numbers with and without the application of DP.

bit confusing presenting both *numerical* and *percent* errors, so I italicize the terms for help readers more easily distinguish which measure is being discussed.

Table 1 shows the number of Unified School Districts and Places that have large errors. They account for a non-trivial share of all Unified School Districts and Places based on the DP Demonstration Produce released by the Census Bureau in August 2022. .   For background, there were about 11,000 Unified School Districts and about 29,000 Places in the 2010 Census which the database reflected in the DP Demonstration Produce released by the Census Bureau in August 2022.

Table 1. Unified School Districts and Places with Large Errors for the Population Age 0 to 4 Due to the Application of Differential Privacy

	Units with Absolute *Numeric* Errors of 25 or more young children		Units with Absolute *Percent* Errors of 25 Percent or More	
	Number	Percent of All Units	Number	Percent of All Units
Unified School Districts	590	5	214	2
Places	564	2	4,209	15

Table 1 shows there are 590 Unified School Districts with absolute *numeric* errors of 25 or more young children and those 590 Unified School Districts are 5 percent of all Unified School Districts. Table 1 shows there 564 Places with absolute *numeric* errors of 25 or more young children and those 564 Places are 2 percent of all Places.

Table 1 shows 214 Unified School Districts with an absolute *percent* error of 25 percent or more young children and they are 2 percent of all United School Districts.

Analysis found 4,209 Places with absolute *percent* errors of 25 percent or more and those 4,209 Places are 15 percent or all places.

If most of the large *percent* errors are based on very small *number*s, they may not too concerning. But the tables below, showing the relationship between large *percent* errors and large *numeric* errors. indicates many units with a high absolute *percent* error also have large absolute *numeric* errors as well.

Table 2 shows there are 590 Unified School Districts with absolute *numeric* errors of 25 or more young children. About 40 percent of these 590 Unified School Districts have absolute *percent* errors of 5 percent or more. In other words, many of the Unified School Districts with large absolute *numeric* errors also have relatively large absolute *percent* errors as well.

Table 2. Distribution of Unified School Districts with Absolute *Numeric* Errors of 25 of More Young Children by Size of Absolute *Percent* Error

Absolute *Percent* Error	Number	Percent of Total
0 to 4.9 Percent	346	59
5 to 9.9 Percent	135	23
10 to 24.9 Percent	97	16
25 Percent or more	12	2
Total	590	100

Table 3 shows the distribution of absolute *numeric* errors for all Unified School Districts with an absolute *percent* error of 25 percent or more. Almost two-thirds (65 percent) of Unified School Districts with absolute *percent error* of 25 percent or more also have absolute *numeric* errors of 5 or more young children. This shows that most of the Unified School Districts with large absolute *percent* errors also have relatively large absolute *numeric* errors.

Table 3. Distribution of Unifed School Districts With Absolute *Percent* Errors of 25 Percent or More by Size of Absolute *Numeric* Error		
Absolute *Numeric* Error	Number	Percent of Total
0 to 4 young children	74	35
5 to 9 young children	59	28
10 to 24 young children	77	36
25 or more young children	4	2
Total	214	101*
* total is not 100% because of rounding.		

Table 4 shows there are 564 Places with absolute p*ercent* errors of 25 Percent or more. More than half (55 percent) of these 564 Places have absolute *numeric* errors of 5 or more young children. In other words, many of the Places with large absolute *numeric e*rrors also have relatively large absolute *percent* errors as well.

Table 4 Distribution of Places with Absolute *Numeric* Errors of 25 or More Young Children by Size of Absolute *Percent* Error		
Absolute *Percent* Error	Number	Percent of Total
0 to 4.9 Percent	256	45
5 to 9.9 Percent	96	17
10 to 24.9 Percent	138	24
25 Percent or more	74	13
Total*	564	99

4

* Total is not 100% because of rounding.

Table 1 shows there are 4,209 Places with an absolute *percent* error of 25 percent or more for young children and they represent 15 percent of all Places. Table 5 also shows the distribution of absolute *numeric* errors for Places with an absolute *percent* error of 25 percent or more. Most (56 percent) had absolute *numeric* errors of 5 or more young children. Table 5 shows there are 74 Places with absolute percent errors of 25 percent or more AND Absolute *numeric* errors of 25 or more young children.

Table 5. Distribution of Places With Absolute *Percent* Errors of 25 Percent or More by Size of Absolute *Numeric* Errors

Absolute *Numeric* Error	Number	Percent of Total
0 to 4 young children	2,267	54
5 to 9 young children	1,222	29
10 to 24 young children	646	15
25 or more young children	74	2
Total	4,209	100

I believe geographic units with large errors will be the biggest problem in terms of the application of DP to the 2020 Census, particularly if such errors are accompanied by changes in funding.   I understand DP is a complicated methodology, but it is not clear to me why the Census Bureau cannot truncate the distribution from which it draws the random numbers used in applying DP, to make sure large errors are not injected into the reported data. If the errors injected by DP could be kept to less than 5 percent or less than 5 people, I believe the application of DP in the 2020 Census would be much more acceptable.

References

O'Hare. W. P. (2022). Analysis of Census Bureau's August 2022 Differential Privacy Demonstration Product: Implications for Data on Young Children, (Sept) . https://secureservercdn.net/198.71.233.229/2hj.858.myftpupload.com/wp-content/uploads/2022/09/Implications-of-Differential-Privacy-for-kids-9-21-2022-FINAL-.pdf

U.S. Census Bureau, (2022) "Demographic and Housing Characteristics File (DHC) Update," Presentation at the Census Bureau's National Advisory Committee meeting September 23, 2022.

5. Werner, Angela, Environmental Health Tracking Section | National Center for Environmental Health | Centers for Disease Control

Hi,

Please see attached for feedback from the Centers for Disease Control and Prevention (CDC) on the August 2022 demonstration dataset. We look forward to further discussion and collaboration on this topic.

Sincerely,

Angie Werner
Science Development Team Lead
Environmental Health Tracking Section | National Center for Environmental Health | Centers for Disease Control

TO:
ron.s.jarmin@census.gov;
christa.d.jones@census.gov;
john.maron.abowd@census.gov;
karen.battle@census.gov;
michael.b.hawes@census.gov;
victoria.a.velkoff@census.gov

CC:
Moyer, Brian (CDC/DDPHSS/NCHS/OD) <qbk2@cdc.gov>;
Werner, Angela (CDC/DDNID/NCEH/DEHSP) <myo6@cdc.gov>;
Bunnell, Rebecca (CDC/DDPHSS/OS/OD) <rrb7@cdc.gov>;
Layden, Jennifer (CDC/DDPHSS/OS/OD) <qbg5@cdc.gov>;
Williamson, G. David (CDC/DDNID/NCEH/OD) <dxw2@cdc.gov>;
Jernigan, Daniel B. (CDC/DDPHSS/OD) <dbj0@cdc.gov>;
Cono, Joanne (CDC/DDPHSS/OS/OD) <bzc6@cdc.gov>

Dear Ron,

I hope this note finds you well.

Thank you and your colleagues for meeting with us to discuss CDC's concerns regarding Differential Privacy and the 2020 Decennial Census. As you recommended, we have re-analyzed our Impact Statements based on the newly available, March 2022 demonstration data. Below, you will find our most recent packet of Impact Statements for your review.

Below is a summary of our key findings, based on the updated Statements:

- Looking at the total population counts, there is general improvement to the data in some areas (Alaska regional areas or rural villages, for example), except for those areas with very small populations. It is worth noting that total population counts may not be the most helpful metric; calculated rates may be preferred, as they are often more responsive to differences in population distributions.
- County-level data show some overall improvement when calculating age-adjusted rates (no stratification by sex or by race/ethnicity). There are still, however, significant differences in rates using the updated data, particularly in counties with smaller populations and when stratifying the age-adjusted rates (for example, the COVID-19, age-adjusted rates by race/ethnicity and by rural/urban areas).
- County-level data remain problematic when estimating age-specific rates, including larger populations of up to 10,000 people.
- Census tract-level data remain problematic when calculating age-adjusted rates. Total population counts may not change significantly, but population changes within individual age groups can significantly impact the overall age-adjusted rate calculations.
- Observations about block-level data:
  - Used by CDC for emergency response purposes to do environmental assessments when working with communities near environmental sites, and other analyses.
  - Noted the Census Bureau's view that block-level data will not be reliable.

- Noted the variation in estimates, regardless of population density.
- CDC will not be able to accurately characterize risks and identify/target vulnerable populations using block-level data.
- Block-level maps will be unreliable, regardless of any aggregation of the block-level data.

We welcome a follow-up discussion and your continued support and collaboration as we explore the best options for using the 2020 Decennial Census data across CDC programs.

Again, thank you very much, and please do not hesitate to reach out.

Warm regards,
Brian

Brian C. Moyer, Ph.D.
Director, National Center for Health Statistics
Centers for Disease Control and Prevention
Phone: (301) 458-4255
Email: brian.moyer@cdc.hhs.gov

CDC differential privacy impact statements

The Differential Privacy Interest Group was convened in early 2020 to discuss issues and questions surrounding the implementation of differential privacy, which is a proposed procedure that injects statistical noise into Census 2020 population and household data products. The move to differential privacy has significant implications for work across all levels of CDC but particularly for the Centers and Divisions that work with human population data. The Interest Group put together impact statements to show the impact that the implementation of differential privacy will have on their work moving forward. This document contains a selection of impact statements.

For further information about the Interest Group or these efforts, please contact Dr. Angela Werner (awerner@cdc.gov).

Table of Contents

CDC COVID-19 Response

Title of project: Assessing the impact of differential privacy on the age-adjusted incidence of COVID-19 at the county level

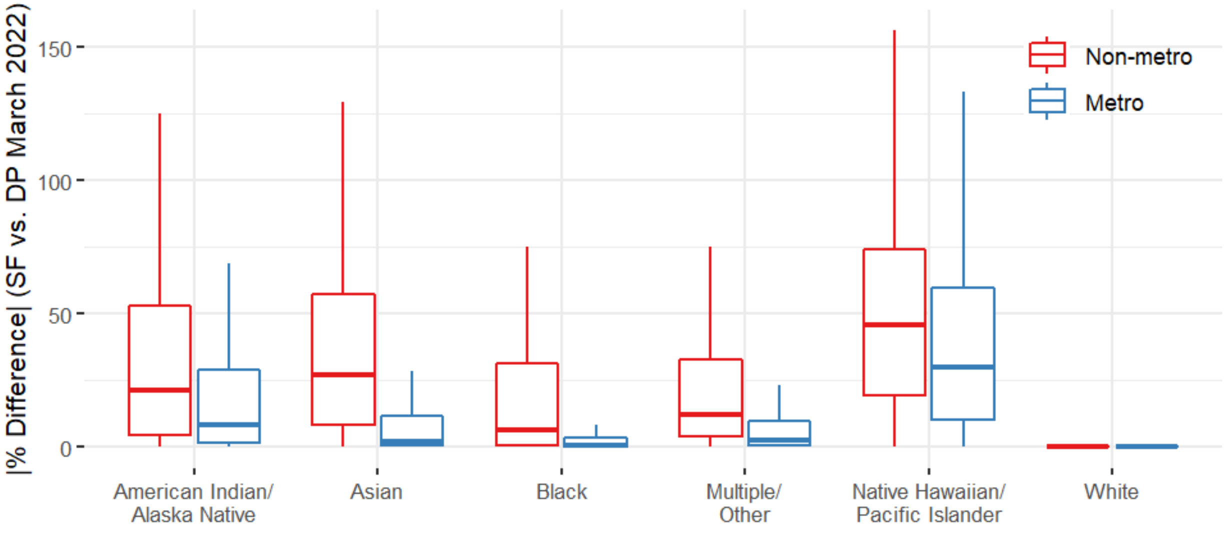
CIO/Division/Program: National Center for Environmental Health/Division of Environmental Health Science and Practice/National Environmental Public Health Tracking Program (facilitated through data collected by the Case Data Section, Data, Analytics, and Visualization Task Force)

Project description: Differential privacy is a statistical adjustment of population counts in public use datasets to protect the privacy of respondents from unauthorized disclosure. When those population counts are used as denominators for computing COVID-19 incidence rates, they may differ from the true values because of the differential privacy adjustment. To facilitate assessment of the expected impact of differential privacy, the Census Bureau released a differentially private version of the 2010 Census data. We used the most recent version of the demonstration dataset (v3-16-2022) to assess the impact of differential privacy on the 2020 age-adjusted incidence of COVID-19 at the county level stratified by race. We calculated age-adjusted incidence as this is a standard measure used in public health, and this allows for comparisons between populations with different age structures.

Methods: The number of COVID-19 cases that occurred in 2020 by county, race, and age were reported to CDC via the National Notifiable Disease Surveillance System (NNDSS), direct data entry (a legacy format used February – October 2020), or by the direct submission of CSV tables by state health departments. Cases were included if case report date, age, race, and county of residence were submitted. Age-adjusted incidence rates were calculated with and without differential privacy by dividing the reported number of COVID-19 cases for 18 separate age groups by the total 2010 population and differentially private 2010 population of each age group, with age standardization completed using the 2000 U.S. Standard Population. The absolute value of the percent difference between the age-adjusted incidence rates generated using the enumerated 2010 population counts and age-adjusted incidence rates generated using the differential privacy demonstration dataset were calculated.

Impact on project: 2020 COVID-19 age-adjusted incidence rates for minority groups were disproportionately affected by the implementation of differential privacy. Age-adjusted incidence rates for American Indian/Alaska Native, Asian, Black, Native Hawaiian/Pacific Islander, and multi-racial/other populations were particularly impacted by differential privacy, especially in non-metropolitan counties (Figure 1). The most significant divergence was observed in Native Hawaiian/Pacific Islander populations with a median difference in age-adjusted incidence rates of 45.7% in non-metropolitan counties and 29.9% in metropolitan counties (Figure 1). Age-adjusted incidence rates among Black populations were highly affected across a broad geographic area, most notably in counties with smaller Black populations (Figure 2).

Societal impact: These results demonstrate the profound impact differential privacy could have on COVID-19 incidence rates by race when differentially private Census 2020 denominators are used to compute those rates. Because racial/ethnic minority populations in counties are disproportionately affected by COVID-19 and other public health threats, the use of differentially private Census 2020 population counts could artificially increase disparities in county-level COVID-19 incidence rates, affecting existing health equity challenges. Differential privacy tends to have greater impacts on smaller populations (e.g., rural populations, minority groups). Because racial or ethnic minority populations are typically smaller populations, the age-adjusted rates for these groups are more sensitive to the effects of differential privacy. The smaller population sizes exacerbated this issue, with especially skewed rates occurring in non-Hispanic American Indian/Alaska Native, non-Hispanic Asian, non-Hispanic Black, non-Hispanic Native Hawaiian/Pacific Islander, and multi-racial/other populations, particularly those populations in rural areas. These large statistical artifacts created by differential privacy can distort the agency's health equity efforts if we are unable to distinguish real increases or decreases in COVID-19 incidence from changes caused by noise injected in the Census population denominators used. In this example, this could result in improperly allocating scarce medical resources during a pandemic and incorrectly targeting or withholding resources for vaccination based on the assumption that increases in COVID-19 incidence are real or unreal.



Source: US Census Bureau 2022; CDC Environmental Public Health Tracking Program

Figure 1: Absolute value of the percent difference between 2020 COVID-19 age-adjusted incidence rates calculated with 2010 enumerated Census population counts and age-adjusted incidence rates calculated with 2010 differential privacy demonstration population counts released in March 2022 (v3-16-2022; most recent version). Red boxes include non-metropolitan counties and blue boxes include metropolitan counties. SF = Summary File 1 data file from Census, which includes data on sex, age, race. DP = Differential Privacy demonstration dataset released in March 2022 to assess the impact of differential privacy, which includes data on sex, age, race.

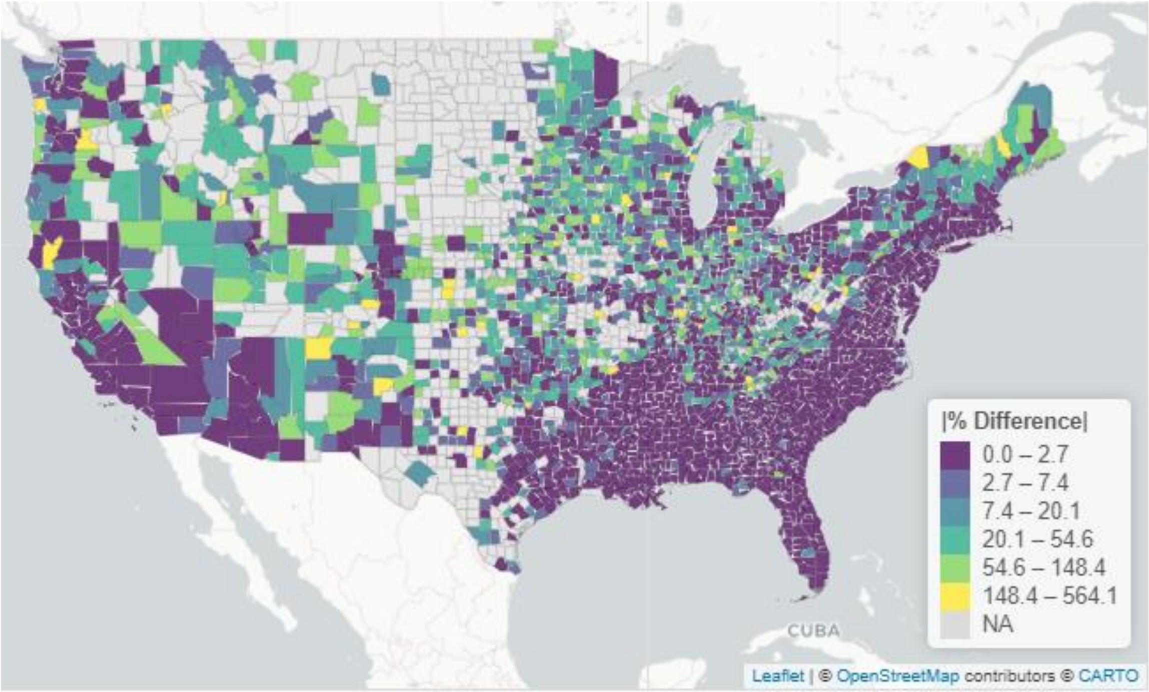


Figure 2: Absolute value of the percent difference between 2020 COVID-19 age-adjusted incidence rates for Black individuals calculated with enumerated 2010 population counts and age-adjusted incidence rates calculated with 2010 differential privacy demonstration population counts released in March 2022 (v3-16-2022; most recent version). NA values occur when no cases were reported among Black individuals in that county during 2020.

National Center for Environmental Health/National Environmental Public Health Tracking Program

Title of project: Assessing the impact of differential privacy on the incidence of health outcomes displayed on the Tracking Network (https://ephtracking.cdc.gov/DataExplorer/) at different geographic levels

CIO/Division/Program: National Center for Environmental Health/Division of Environmental Health Science and Practice/National Environmental Public Health Tracking Program

Project description: In order to facilitate assessment of the impact of differential privacy, the Census Bureau released a differentially private version (v3-16-2022) of the 2010 Census data. We used this demonstration dataset to assess the implications of differential privacy on age-adjusted rates of asthma emergency department (ED) visits and acute myocardial infarction (AMI) hospitalization at the county and census tract levels.

Methods: County- and census tract-level counts of asthma ED visits and AMI hospitalizations were acquired from recipients of the Environmental Public Health Tracking Program. In total, asthma ED data were acquired from 30 states at the county level and 6 states at the census tract level. AMI hospitalization data were acquired from 31 states at the county level and 7 states at the census tract level. Age-adjusted rates were calculated with and without differential privacy by dividing the reported number of asthma ED visits and AMI hospitalizations for 18 separate age groups by the total 2010 population and differentially private 2010 population of each age group, with age standardization completed using the 2000 U.S. Standard Population.

Impact on project: Differential privacy had minimal effects on the estimated rate of asthma ED visits at the county level (Fig. 1a). At the census tract level, changes in the rate of asthma ED visits were **generally less than 2-fold**, though differential privacy had significant effects in several census tracts, including one in which the rate of asthma ED visits **increased over 400-fold** (Fig. 1b). Only minor changes in AMI hospitalization rates were detected at the county level (Fig. 2a), though changes in the rate of hospitalizations at census tract level **routinely exceeded 5%** (Fig. 2b). While the total population count typically did not change substantially as a result of differential privacy, age-adjusted rates were sensitive to population changes within individual age groups at the census tract level.

Societal impact: Changes in population counts due to differential privacy could result in significantly overestimated (or significantly underestimated) rates, particularly at finer spatial resolutions such as census tract level. Small population sizes tended to exacerbate this issue with especially skewed rates occurring at the census tract level. This is particularly important as the Tracking Program moves to displaying and disseminating sub-county data.

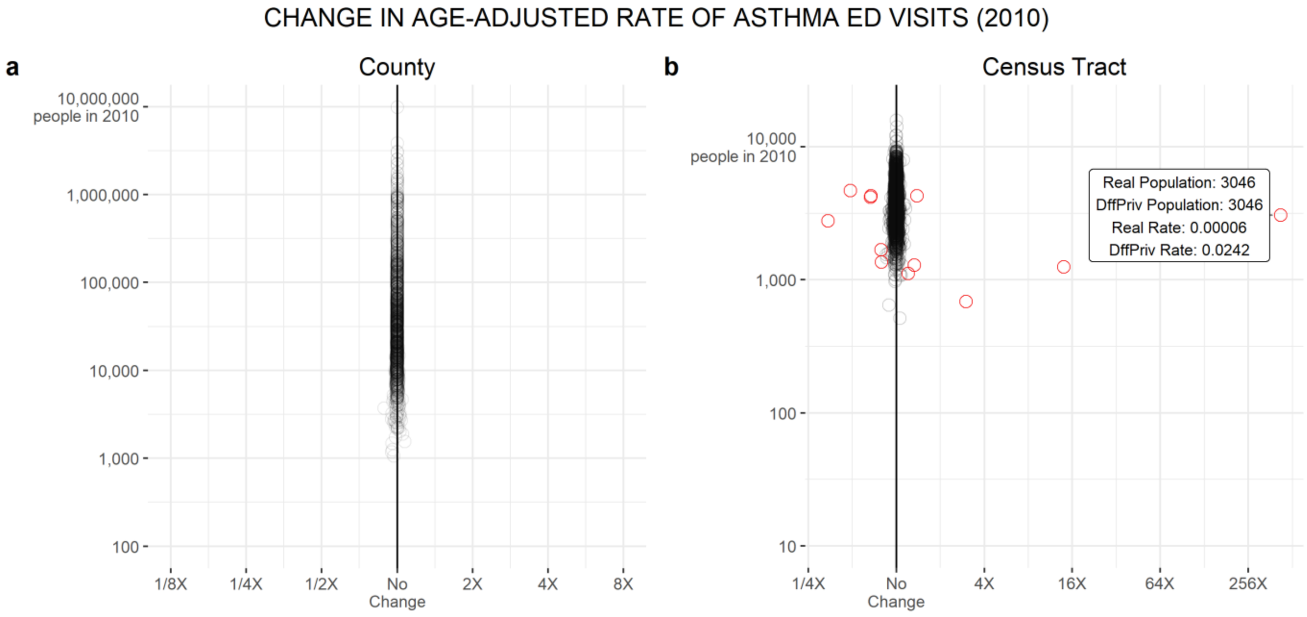


Figure 1: Change in the age-adjusted rate of asthma emergency department visits in 2010 due differential privacy population adjustments at the a) county (30 states) and b) census tract level (6 states). Counties and census tracts with at least a 20% change from the true rate are in red.

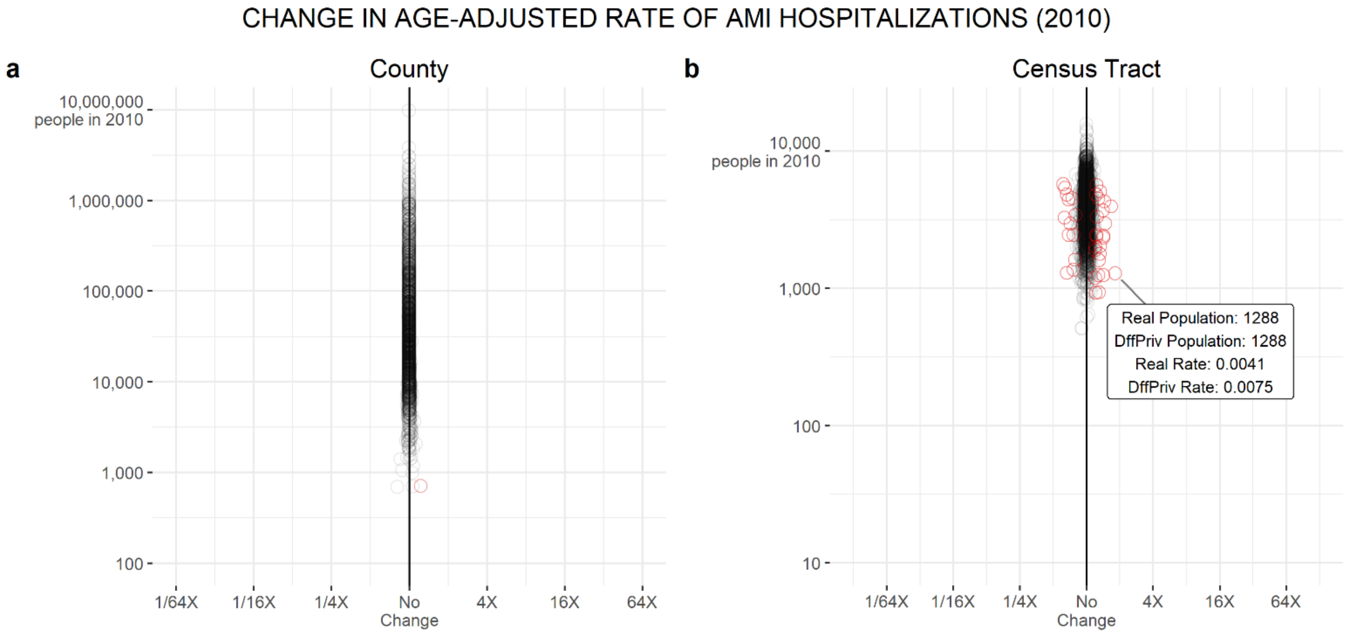


Figure 2: Change in the age-adjusted rate of acute myocardial infarction (AMI) hospitalization in 2010 due differential privacy population adjustments at the a) county (31 states) and b) census tract level (7 states). Counties and census tracts with at least a 20% change from the true rate are in red.

National Center for Chronic Disease Prevention and Health Promotion/Division for Heart Disease and Stroke Prevention

Title of project: Assessing the impact of differential privacy on estimated county-level heart disease mortality overall and by sub-group

CIO/Division/Program: National Center for Chronic Disease Prevention and Health Promotion/Division for Heart Disease and Stroke Prevention

Project description: To facilitate the assessment of the potential impact of differential privacy, the Census Bureau released a differentially private version of the 2010 Census population data. We used the most recent version (v3-16-2022) of the demonstration dataset to assess the implications of differential privacy on estimated county-level death rates for heart disease, the nation's leading cause of death.

Method: We obtained county-level heart disease death counts for the year 2010 from the National Vital Statistics System in the National Center for Health Statistics (NCHS). With these death counts, we then estimated county-level rates using two sets of denominators: (1) bridged-race populations provided by NCHS and (2) the U.S. Census Bureau's differentially private populations. To generate these estimates, we used a Bayesian spatiotemporal conditional autoregressive model that has been used extensively to examine spatiotemporal trends in cardiovascular disease death rates. Briefly, this model estimates more precise, reliable rates by incorporating correlation across space and demographic group, even in the presence of small death counts and small populations.

With this model and each population dataset, we estimated two sets of county-level rates for (1) the entire population (i.e., overall rates), and (2) stratified by both 10-year age groups and sex, resulting in four sets of rate estimates. The overall death rates were age-standardized to the 2010 U.S. population using 10-year age groups. These two sets of rates represent scenarios based on higher death counts and populations (overall rates) and based on smaller death counts and populations (rates by age group and sex). We calculated the percent change between the rates generated using the NCHS populations and the Census differential privacy populations.

Impact on project: For the overall age-standardized rates, only 0.3% of rates had more than a 20% difference between the rates estimated using differential privacy and NCHS populations (Figure 1). However, for the rates stratified by age group and sex, almost half (43.3%) of rates had more than a 20% difference between the rates estimated using differential privacy and NCHS populations. For both the overall and stratified rates, 98.0% of rates with more than a 20% difference occurred in populations of less than 10,000 people. Although some rates estimated using the differential privacy populations were lower than those estimated using the NCHS population, higher estimates were more common.

Societal impact: This analysis shows that changes in population counts resulting from the differential privacy algorithm could lead to large differences in the estimates of heart disease death rates, especially for populations less than 10,000 people. This change would severely hamper the ability to report and intervene upon heart disease in small populations, such as rural counties and among for some racial/ethnic groups. More specifically, differential privacy could impact the ability to report on county-level death rates for the total population in rural areas and for subpopulations (e.g., by race and Hispanic ethnicity, age group) in many counties across the country. This process would especially hamper surveillance of cardiovascular disease mortality within racial and Hispanic ethnic groups, many of which have higher mortality, and for younger adults, which have low but increasing cardiovascular disease mortality. Higher estimates in these smaller populations would mask the places and groups with truly high estimates. These potential problems with surveillance become magnified as these estimates are disseminated to state and local health departments for their program planning and resource allocation.

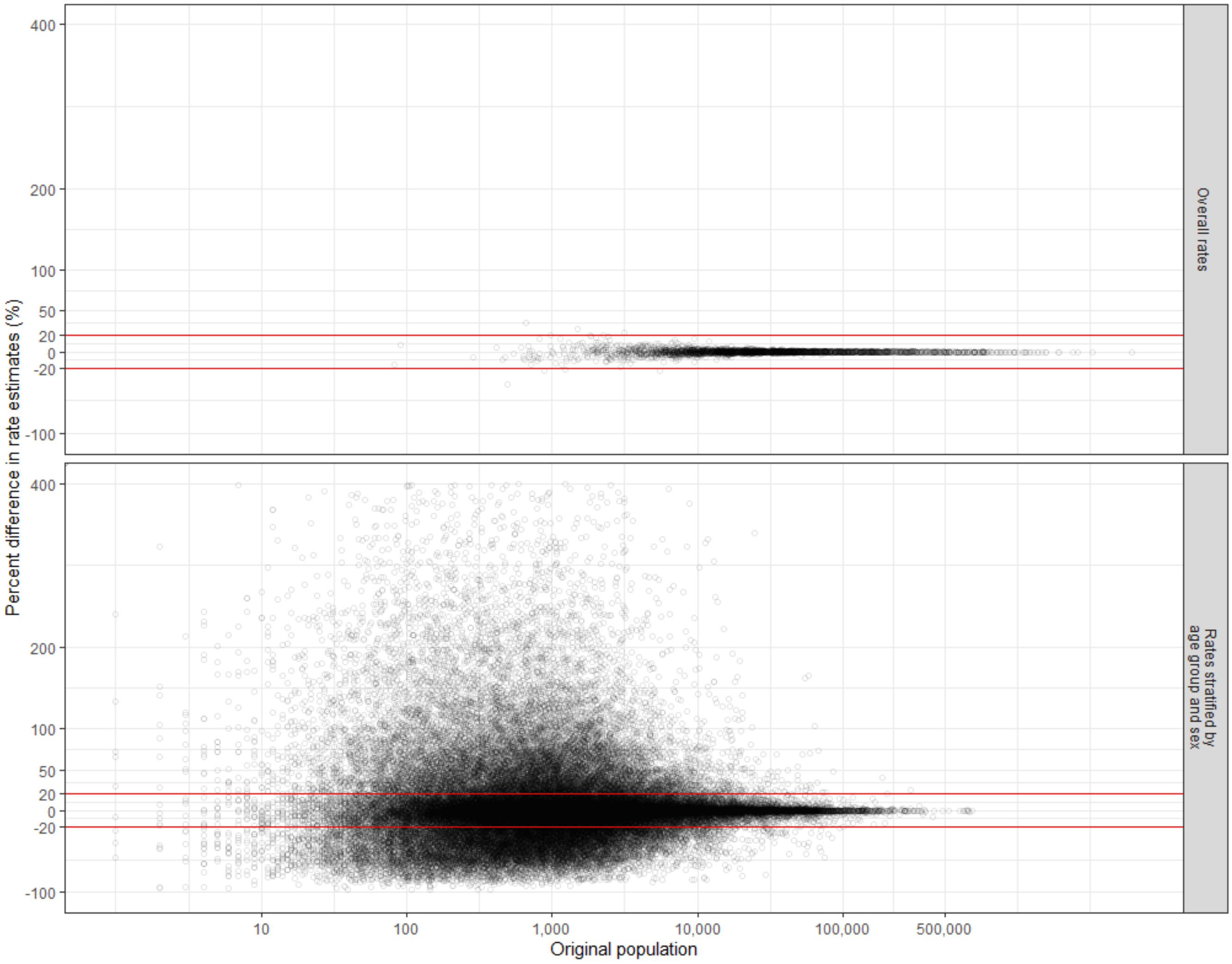


Figure 1: Percent difference in age-adjusted county-level heart disease death rates between NCHS populations and differential privacy population estimates. A positive difference indicates that the rate estimated using differential privacy populations was a higher value. Outliers have been truncated from this figure.

National Center for Emerging and Zoonotic Infectious Diseases/Arctic Investigations Program

Title of project: Assessing the impact of differential privacy on the accuracy of estimates of infectious disease burden within Alaska and among Alaska Native peoples

CIO/Division/Program: National Center for Emerging and Zoonotic Infectious Diseases/Division of Preparedness and Emerging Infections/Arctic Investigations Program

Project description: The Arctic Investigations Program (AIP) is an infectious disease field station located on the campus of the Alaska Native Tribal Health Consortium in Anchorage, Alaska. Their mission is the prevention of infectious diseases in the arctic and sub-arctic with a special emphasis on diseases of high incidence among Alaska Native people. AIP conducts infectious disease surveillance on many pathogens and here we describe the potential impact of the U.S. Census differential privacy on accuracy of estimates of disease burden and disease disparity within the state of Alaska.

Methods: We primarily describe the breadth of AIP's disease surveillance for infectious disease and the importance of U.S. census denominators in our ability to accurately examine disease disparities and social determinants of health within the state of Alaska. Population estimates from the 2010 census were compared to the original differential privacy (DP) release, the May 27, 2020 release (PPMF), and the most recent March 16, 2022 DP release (DP2).

Impact on program: In order to assess racial health disparities and underlying social determinants of health within Alaska, AIP routinely calculates incidence and prevalence for a wide range of infectious diseases and their associated medical consequences. They include but are not limited to: invasive bacterial diseases (*Streptococcus pneumoniae*, *Haemophilus influenzae*, Group A *streptococcus* and *Helicobacter pylori*), hepatitis A, B, and C, respiratory syncytial virus, influenza, and a variety of zoonotic infections, and gastric and liver cancer. Alaska is a state with a 2019 population of approximately 735,000 persons, 20% of whom are Alaska Native peoples. Assessing racial disparities in disease burden requires age adjustment. To examine underlying determinants of health, the AIP program often must further stratify by geographical subregions (regional populations range from a low of 6,000-400,000 persons) or rural village (range from ~100 – 1,000 persons). For example, community-level disease burden estimates have been used to evaluate the relationship between the presence of running water in the homes of rural village communities and the rates of infectious diseases (1-3) and as part of a recent investigation into severe invasive *Haemophilus influenzae* disease in a single Alaskan community (4). Using data from the first differential privacy release (DP) and the May 2020 release (PPMF) would result in substantial percent errors in the denominator for villages with a population < 500 persons, resulting in imprecise disease burden estimates (Table 1). The most recent March 16, 2022 release (DP2) has reduced the errors considerably with only villages with a population < 100 persons in Alaska continuing to lack precision (Table 1). When broken down by gender, age, and Alaska Native race at the regional level, denominator accuracy diminishes sharply between our largest region and one of our smaller regions (Figure 1) in the first two releases. However, the amount of error, at the regional level, has substantially improved between the PPMF release and the DP2 release (Figure 2). Alaska is small enough in population, that assessments within the smaller regions and at the village level (particularly villages < 100 persons) will still continue to be impacted by differential privacy, but to a smaller degree with the DP2 release.

Societal impact: The AIP's ability to accurately estimate disease burden within the State of Alaska will be impacted by the proposed differential privacy release. Loss of precision in population denominators could impact public health programs within Alaska aimed at mitigating disease disparities at the village and regional level. The small population numbers within the state of Alaska make this unique region of the United States particularly vulnerable to the impacts of differential privacy. Disease burden estimates are used to inform public health prevention programs within the State of Alaska. Examples of prevention efforts include vaccine development, measurement of vaccine impact, enhanced medical screening and highlighting the negative health consequences associated with lack of running water in some rural Alaska homes.

Table 1. Mean absolute percent error (MAPE) for differential privacy release according to size of Alaska community. DP1 is the original differential privacy release from the U.S. Census and PPMF and DP2 are the subsequent release using differential privacy from May 27, 2020 and March 16, 2022, respectively.

Places and CDPs				
Population	MAPE-DP (Original Release)	MAPE-PPMR (May 2020 Release)	MAPE-DP2 (March 2022 Release)	N
1-100	**131%**	**92%**	**7%**	**107**
101-500	**15%**	**13%**	**1%**	**127**
501-1,000	**6%**	**6%**	**0.2%**	**53**
1,001-5,000	**2%**	**2%**	**0.2%**	**43**
5,001-10,000	**1%**	**1%**	**0%**	**16**
10,000+	**1%**	**0%**	**0%**	**6**

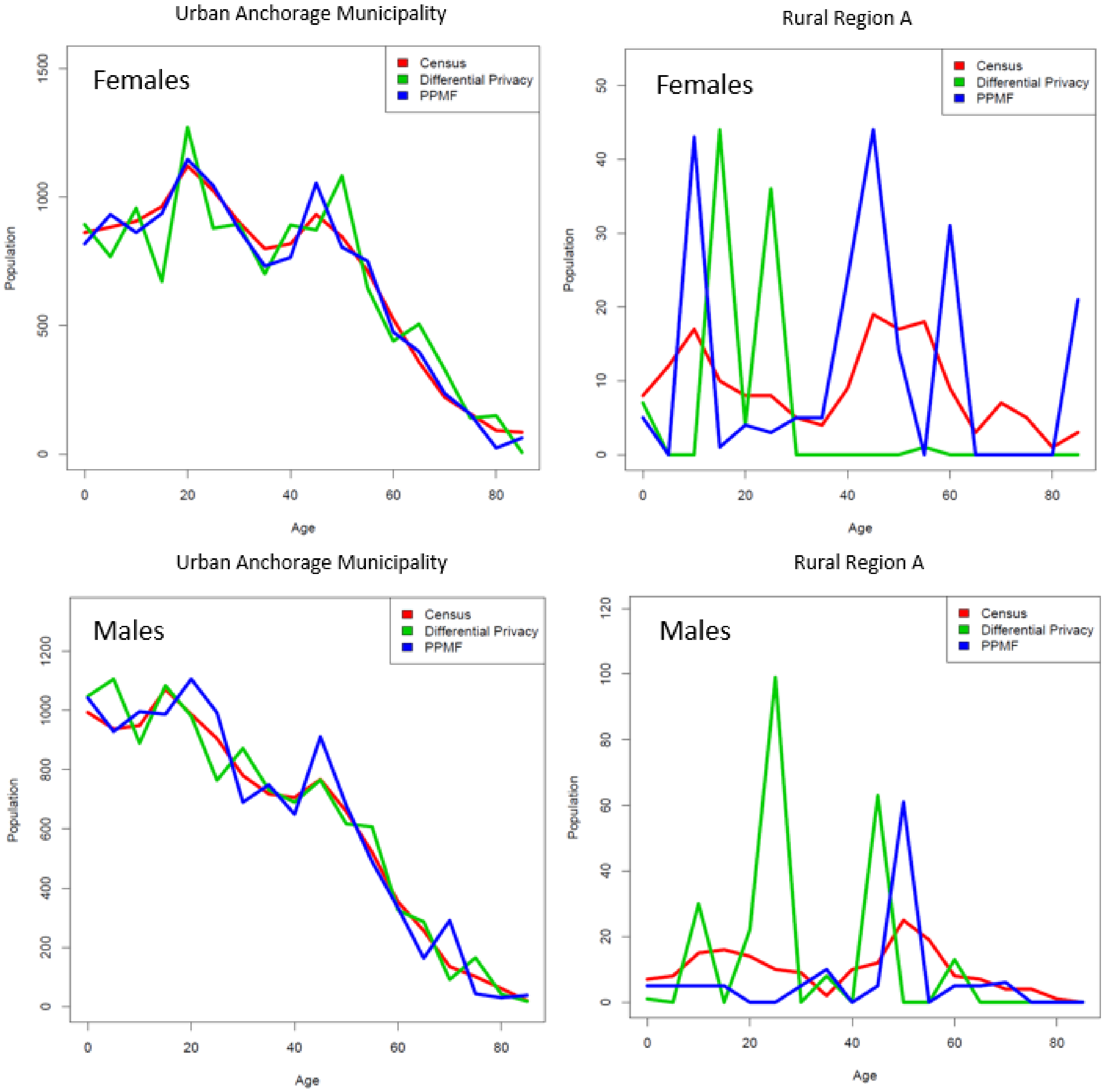


Figure 1. Population estimates for 2 regions of Alaska according to the U.S. Census original release, differential privacy release and the May 27, 2020 release (PPMF). Estimates are for Alaska Native persons and broken down by gender and age.

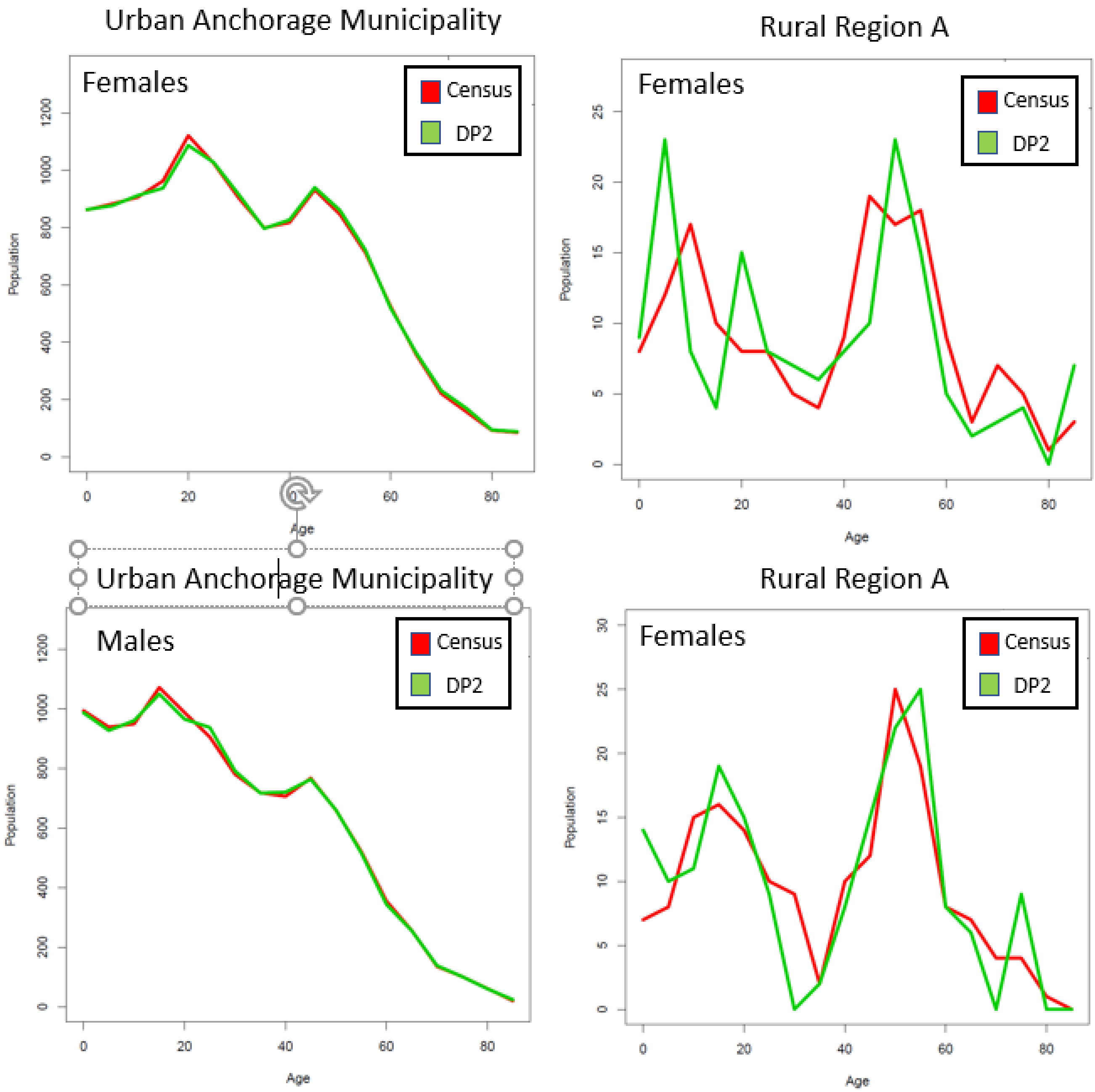


Figure 2. Population estimates for 2 regions of Alaska according to the U.S. Census original release, and the most recent March 16, 2022 release (DP2). Estimates are for Alaska Native persons and broken down by gender and age.

<u>Citations</u>:

1) Mosites, E, B Lefferts, S. Seeman, et al. Community water service and incidence of respiratory, skin, and gastrointestinal infections in rural Alaska, 2013-2015. Int J Hyg Environ Health, April 2020, ePUB ahead of print.

2) Thomas, TK, T Ritter, D Bruden, M Bruce, K Byrd, R Goldberg, K Hickel, J Smith, and T Hennessy. Impact of Providing In-home Water Service on the Rates of Infectious Diseases: Results from Four Communities in Western Alaska, J of Water and Health, 2016, 14(1): 132-141.

3) Hennessy, TW, T Ritter, RC Holman, DL Bruden, KL Yorita, LR Bulkow, JE Cheek, RJ Singleton, and J Smith. The Relationsip Between In-Home Water Service and the Risk of Respiratory Tract, Skin, and Gastrointestinal Tract Infections Among Rural Alaska Natives, American Journal of Public Health, 2008 98(11): 2072-8, November.

4) Nolen, L, A Tiffany, C DeByle, D Bruden, G Thompson, A Reasonover, D Hurlburt, et al. Haemophilus influenzae serotype a (Hia) carriage in a small Alaska community after a cluster of invasive Hia disease, 2018. Clin Inf Dis, June 2020 ePUB ahead of print.

Agency for Toxic Substances and Disease Registry/Geospatial Research, Analysis, and Services Program

Title of project: Assessing the impact of differential privacy on ATSDR's Site Introductory Maps at the block level

CIO/Division/Program: Agency for Toxic Substances and Disease Registry (ATSDR)/ Office of Innovation and Analytics/ Data and Analysis Section/Geospatial Research, Analysis, and Services Program (GRASP)

Project description: The Census Bureau plans to use differential privacy (DP) for the 2020 Decennial Census as a method to reduce disclosure avoidance. The DP method introduces random noise into its 100% count data to ensure privacy. While the Census contends that overall population counts will be preserved, it is likely that there will be adverse impacts to accuracy, particularly in smaller geographic enumeration units. GRASP relies on Census data heavily to support demographic analysis in its application of geospatial science to public health. GRASP used the Census's differential privacy demonstration data to evaluate the potential impact on one of its commonly-produced products—the Introductory Map (aka IntroMap) Series. The IntroMap Series enables scientists to explore and understand the complex and interrelated environmental, sociodemographic, topographical, and cultural conditions that are associated with chemical exposure and the development of acute and chronic conditions. One core map, the General Site Profile Map, depicts the environmental waste site of interest (e.g., EPA National Priority List site), along with any airport, industrial, military, or park land uses. It also provides community demographic and housing statistics at the U.S. Census block level. It is published in many of the public health assessments and health consultations conducted by the Agency for Toxic Substances and Disease Registry (ATSDR).

Methods: GRASP downloaded block-level Census DP demonstration data for Florida and New Jersey from IPUMS National Historical Geographic Information System (NHGIS) and filtered the data for the racial, ethnic, and age fields required in the IntroMap (Tables 1 and 2). Using the 2010 published data and DP 2010 demonstration data downloaded March 2022 (v3-16-2022), GRASP generated demographics and maps for communities near two environmental sites, one located in a densely populated area, and one located in a low-density population area. GRASP selected the Keegan Landfill site in Kearny (Hudson County), New Jersey (Figure 1) to represent a higher density community. GRASP selected the Arkla Terra Property site in Thonotosassa (Hillsborough County), Florida (Figure 2) to represent a less dense population area. GRASP reviewed the population estimates and maps to identified variations between the two Census data sources.

GRASP generated a General Site Profile Map for each site using the GRASP ArcGIS Arc Toolbox (python programming language). The tool generated a map and demographic summaries using the 2010 decennial census variables and one American Community Survey (ACS) variable at the block level. GRASP used the Arc Toolbox tools to enumerate the population based on the same variables from the Census demonstration data. All variables except the age of housing were available in the demonstration data. The Toolbox tool uses the area proportion technique to estimate population values found in blocks that are not completely within the designated buffer area from the site boundary. In those blocks, the population estimate is calculated based on the proportion of the area of the block that is included within the specified distance. Worth noting is that the accuracy of the area proportion technique is directly related to the geospatial granularity of the data so that blocks would produce more dependable estimates than would, for examples, block groups or tracts. One mile was the chosen distance to generate the population estimates. See the buffer rings (¼, ½, 1, 2-mile) in each of the figures presented below.

Impact on project: Differences in demographics statistics varied between the 2010 data published by the Census and the 2010 demonstration data for both the Keegan Landfill site and the Arkla Terra site. As expected, the blocks with lower population counts showed much higher percentages in changes for each group. Tables 1 and 2 contain the population estimates that are found on the IntroMaps for Keegan Landfill and Arkla Terra respectively.

Societal impact: This exercise demonstrates there are differences in population estimates derived from the demonstration data made available by the U.S. Census at the block level. The differences are evident in both tabular and map products. Two example sites are not a representative sample; however, these examples give insight to the potential impact of population data affected by differential privacy methodology. Variations in population estimates were found for both sites regardless of population density. In this exploration, the variations were greater and more noticeable for the site located in a less densely populated area in Florida. The size of the sites may also impact the effect of differential privacy in the population estimates. Differences in population estimates surrounding larger more expansive sites may be more minimized than differences in population estimates surrounding smaller sites. However, the size of the site has no bearing on any differences in the demographic information depicted within the individual blocks on the maps.

Environmental health scientists rely on population estimates to better identify and assess risk to general and sensitive populations living near environmental hazards. Any inaccuracies produced from the application of data affected by differential privacy have the potential to result in a mischaracterization of risks due to faulty understanding of the placement and/or misplacement of sensitive population groups. Exposure assessments, which determine whether people in a community have been exposed to a hazardous substance, require accurate information about where people live to provide more powerful results. An underestimation of population in proximity to chemical contamination may inadvertently exempt affected areas from consideration for these types of vital environmental assessments. As the disclosure avoidance system evolves, the overall population estimates may improve, however mapping products at a more granular geographic level would be suspected to be invalid particularly in less densely populated communities. Further evaluation as resources become available can help further the evaluation of the potential impact of differential privacy. The Census Bureau should provide clear guidance to its end users on how these data anomalies and variations should be communicated in products derived using their published data.

Table 1. for Keegan Landfill, shows a range of 0.1%-36.4% differences in demographic estimates. One population estimate based on the 2010 demonstration data was greater than 20% different from the 2010 published data. The measure with the greatest variation was Native Hawaiian & Other Pacific Islander (36.4% greater). Two measures which fell between 5% and 20% different were American Indian & Alaska Native (14.9% fewer) and Black (7.6% fewer).

Table 1. Keegan Landfill 2010 vs 2010 Demo with DP (v3-16-2022)

Demographic Statistics Within One Mile of Site Boundary			
Measure	**2010**	**Demo**	**Change**
Total Population	35,433	35,347	-0.2%
White Alone	25,333	25,350	+0.1%
Black Alone	847	782	-7.6%
Am. Indian & Alaska Native Alone	161	137	-14.9%
Asian Alone	2,359	2,375	+0.7%
Native Hawaiian & Other Pacific Islander Alone	11	15	+36.4%
Some Other Race Alone	5,359	5,339	-0.4%
Two or More Races	1,367	1,362	-0.4%
Hispanic or Latino (of any race)	15,357	15,300	-0.4%
Children Aged 6 and Younger	3,073	3,056	-0.6%
Adults Aged 65 and Older	3,678	3,702	+0.7%
Females Aged 15 to 44	7,794	7,771	-0.3%
Housing Units	13,116	N/A	N/A

Table 2. for the Arkla Terra Property, shows a range of 1.3%-100% differences in demographic estimates. Two population estimates based on the 2010 demonstration data were greater than 20% different from the 2010 published data. The top measure with the greatest variation was Native Hawaiian & Other Pacific Islander (100% fewer). The second highest was American Indian & Alaska Native (30.4% fewer). Four measures which fell between 5% and 20% different were Some Other Race Alone (15.1% fewer), Hispanic or Latino (9.7% fewer), Two or more Races (8.5% fewer), and Black Alone (5.4% more).

Table 2. Arkla Terra Property 2010 vs 2010 Demo with DP (v3-16-2022)

Demographic Statistics Within One Mile of Site Boundary			
Measure	**2010**	**Demo**	**Change**
Total Population	3,007	2,968	-1.3%
White Alone	2,404	2,363	-1.7%
Black Alone	411	433	+5.4%
Am. Indian & Alaska Native Alone	23	16	-30.4%
Asian Alone	33	32	-3.0%
Native Hawaiian & Other Pacific Islander Alone	1	0	-100.0%
Some Other Race Alone	53	45	-15.1%
Two or More Races	82	75	-8.5%
Hispanic or Latino (of any race)	298	269	-9.7%
Children Aged 6 and Younger	255	262	+2.7%
Adults Aged 65 and Older	436	450	+3.2%
Females Aged 15 to 44	527	546	+3.6%
Housing Units	1,411	N/A	N/A

Figures 1 (Keegan Landfill) and 2 (Arkla Terra) represent the maps drawn from information developed from the Census's published data and demonstration data for each of the sites. The maps developed based upon the 2010 demonstration data (v3-16-2022) highlight variations in all measures of sensitive population groups depicted (Children Ages 6 and Younger, Adults 65 Years and Older, and Females Aged from 15 to 44) surrounding these sites. The variations in Keegan Landfill area populations are not significant enough to change the classification assignments for the blocks within one mile of the sites. On the other hand, several variations do change the classification assignments in the blocks within one mile of the Arkla Terra site. These variations would suggest the maps are not valid at the block level with the disclosure avoidance system applied in lower population areas.

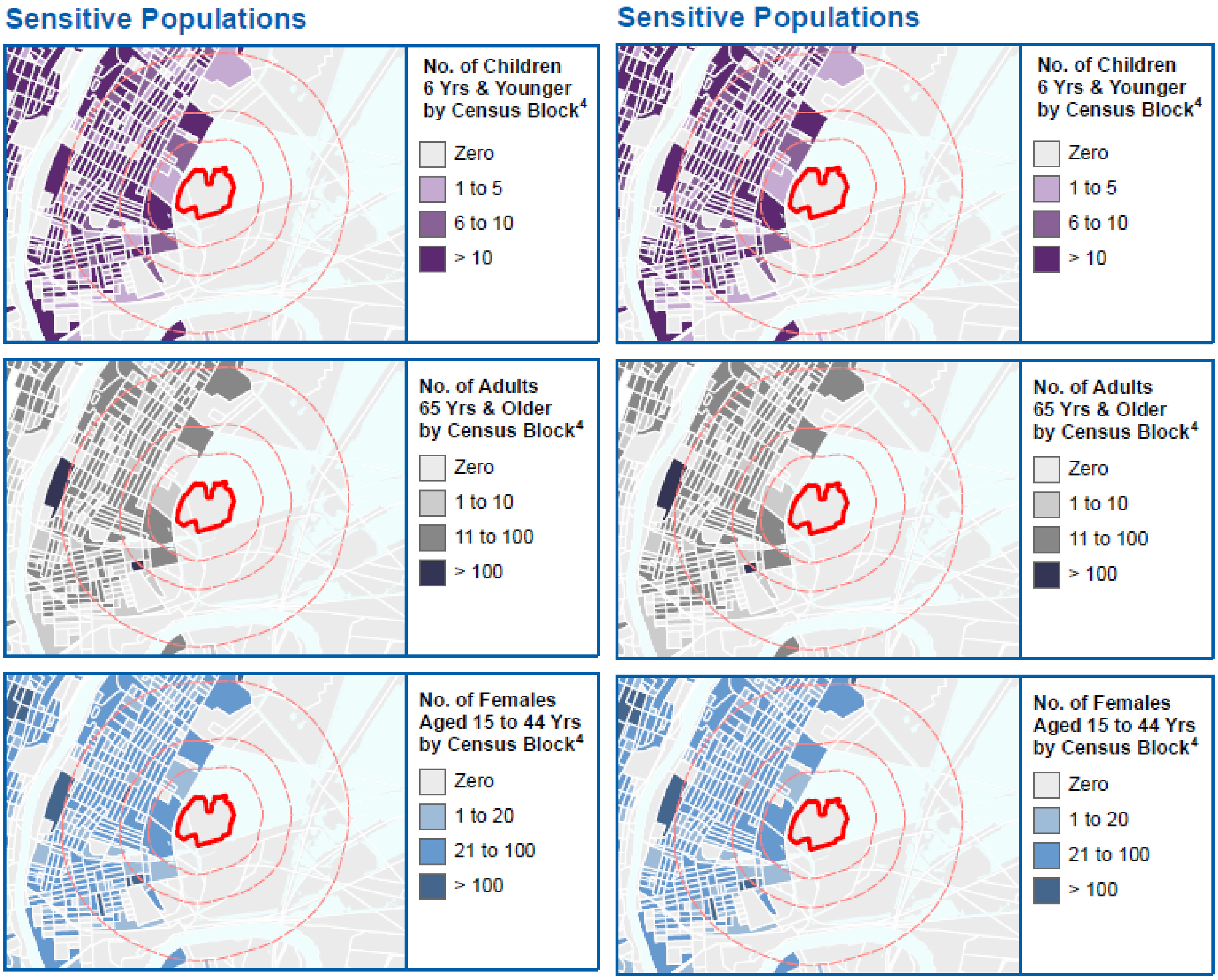

Sensitive Populations
No. of Children 6 Yrs & Younger by Census Block4
Zero
1 to 5
6 to 10
> 10
No. of Adults 65 Yrs & Older by Census Block4
Zero
1 to 10
11 to 100
> 100
No. of Females Aged 15 to 44 Yrs by Census Block4
Zero
1 to 20
21 to 100
> 100
Sensitive Populations
No. of Children 6 Yrs & Younger by Census Block4
Zero
1 to 5
6 to 10
> 10
No. of Adults 65 Yrs & Older by Census Block4
Zero
1 to 10
11 to 100
> 100
No. of Females Aged 15 to 44 Yrs by Census Block4
Zero
1 to 20
21 to 100
> 100

Figure 1. Keegan Sensitive Population Derived Using 2010 Published Data (Left) and 2010 Demo Data (Right)

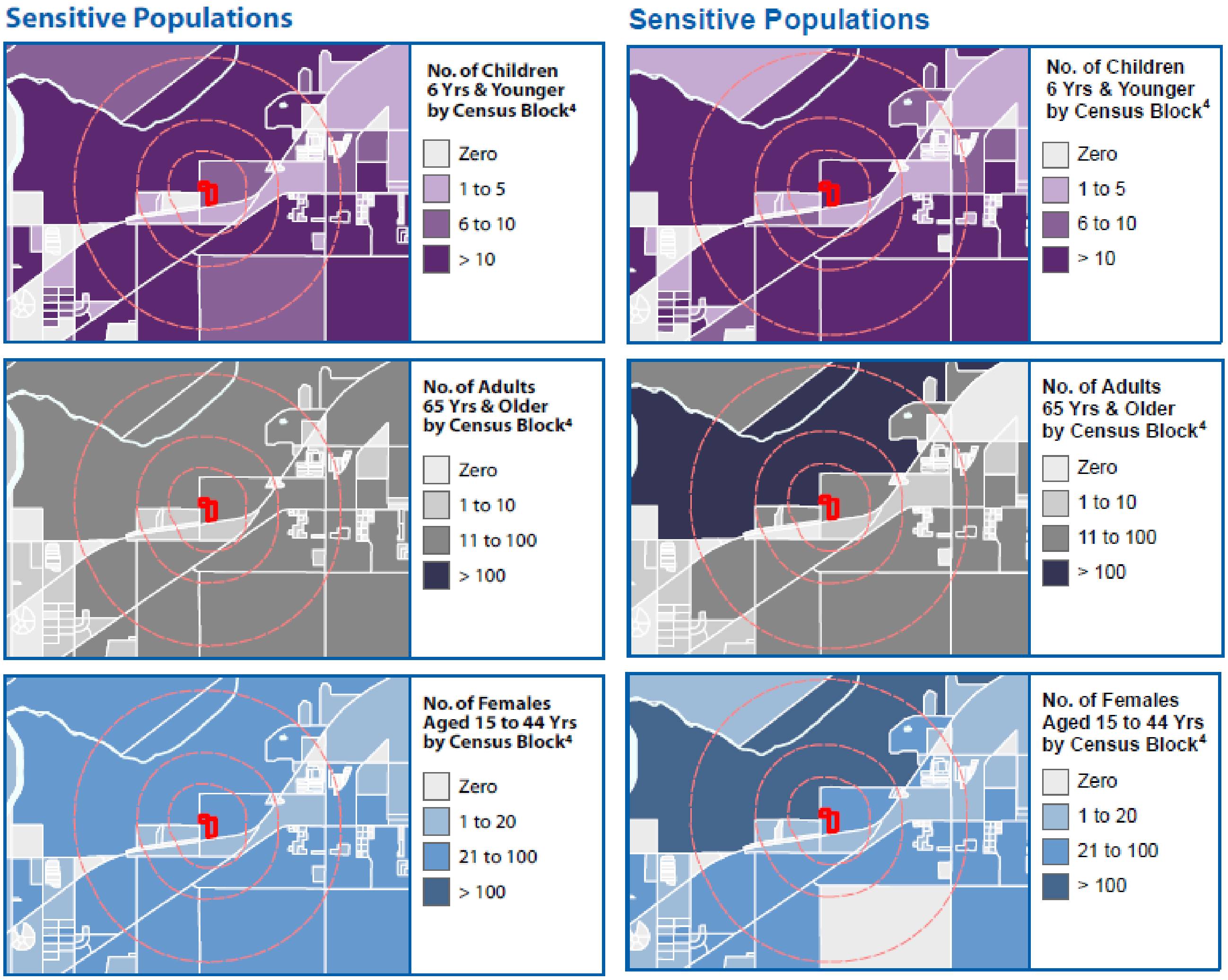

Sensitive Populations
No. of Children 6 Yrs & Younger by Census Block4
Zero
1 to 5
6 to 10
> 10
No. of Adults 65 Yrs & Older by Census Block4
Zero
1 to 10
11 to 100
> 100
No. of Females Aged 15 to 44 Yrs by Census Block4
Zero
1 to 20
21 to 100
> 100
Sensitive Populations
No. of Children 6 Yrs & Younger by Census Block4
Zero
1 to 5
6 to 10
> 10
No. of Adults 65 Yrs & Older by Census Block4
Zero
1 to 10
11 to 100
> 100
No. of Females Aged 15 to 44 Yrs by Census Block4
Zero
1 to 20
21 to 100
> 100

Figure 2. Arkla Terra Sensitive Populations Derived Using 2010 Published Data (Left) and 2010 Demo Data (Right)

6. Eric Guthrie, Minnesota State Demographic Center

Please find the attached feedback on the demonstration data released on August 25, 2022.

I do not have any concern this feedback being shared in any forum where it would be useful. Similarly, I do not have a concern with any statements in the attached letter being attributed to me.

Thank you,

Eric A. Guthrie, PhD
Senior Demographer

MINNESOTA STATE
DEMOGRAPHIC CENTER
DEPARTMENT OF ADMINISTRATION

September 26, 2022

U.S. Census Bureau
4600 Silver Hill Road
Suitland, MD 20746
2020DAS@census.gov

To Whom It May Concern:

Thank you for the opportunity to review these demonstration data and for the ability to provide feedback. I appreciate the ability to provide my insights and how I think it will affect some uses in the State of Minnesota.

I have constrained my review to a single use case, that of making county-level, cohort component population projections. I am looking at the county level as it is a common geographic level for population projections, and it is one that is regularly published by a variety of governmental and non-governmental organizations throughout the nation. According to a recent survey (2021) of the Federal-State cooperative for Population Estimates / Federal-State Cooperative for Population Projections (FSCPE/FSCPP) membership, 45 states and the District of Columbia used the cohort component method for their most recent population projections. Some states produce projections on a regular cycle, and some on an ad hoc basis, but the fact remains that the preferred method is one that requires an accurate population structure by sex to achieve validity. This figure does not include the projections that are produced for and by other governmental and non-governmental agencies.

A first step in developing a projection series is to establish a base population and use that base population to develop specific rates (fertility, mortality, migration, etc.) Those rates will be applied to the base population as they are stepped through the projection period. Given the short nature of the comment period, I could not produce a projection series with the DP and SF1 data to see how they would be different, but I have considered that process as I reviewed at the data from table P12—Age by Sex. I would have used table P12 as a base population were I to have produced projections for this review. To facilitate the review, I have aggregated the age groups below 85 plus into five-year bins to make the matrix more uniform for my analysis.

Sensitivity to error in population projections will vary greatly depending on the size of the area being projected. Counties with comparatively small populations are particularly vulnerable to problems because of the issue of applying rates to relatively small numbers. Given that known issue, I am very concerned that I see a large number of unacceptable errors in the five-year age/sex groups for many counties in Minnesota. I have defined an unacceptable error as being greater than or equal to plus or minus three percent in a particular age bin. In reviewing the data in table P12, after adjusting the bins to be uniform for ages under 85, I have found that only 33 of Minnesota's 87 counties had age structures that did not include at least one bin with an error that is unacceptable.

When looking at counties with a total population of less than 10,000, these data are completely unusable for the purpose of cohort-component projections. Several counties of this size have unacceptable errors in over half of the 36 age-sex bins. Results are shown in Table 1 below. The 19 counties in Table 1 are 22 percent of all the counties in the State of Minnesota. This is especially concerning as the decennial census is the one opportunity for counties of this size to get usable data for their myriad uses as the margins of error in the American Community Survey (ACS) data are often too large to use for most purposes.

Table 1 – Count of Unacceptable Errors in Counties Under 10,000 Residents

County	SF1 Total Population	Unacceptable Age/Sex Bin Error Count
Big Stone County	5,269	9
Clearwater County	8,695	9
Cook County	5,176	20
Grant County	6,018	10
Kittson County	4,552	14
Lac qui Parle County	7,259	12
Lake of the Woods County	4,045	18
Lincoln County	5,896	13
Mahnomen County	5,413	19
Marshall County	9,439	7
Murray County	8,725	7
Norman County	6,852	9
Pipestone County	9,596	7
Red Lake County	4,089	23
Rock County	9,687	6
Stevens County	9,726	6
Swift County	9,783	6
Traverse County	3,558	24
Wilkin County	6,576	16

The potential damage to the ability to produce accurate population projections is not limited to counties in the state below 10,000 residents. When we look at counties that have between 10,000 and 30,000 residents, there are still many counties that have unacceptably large errors in the age-sex data provided in table P12. In that size class of counties, 25 have errors that are larger than three percent in at least one (usually more) age/sex bin. Often there are differences in the 85 plus bin, which presents concrete planning issues, however what is of even more concern are the numerous counties in this size class that have unacceptable errors in female bins between ages 15 and 45. Those are crucial data as they not only have mortality applied, but they are what the fertility rates are applied to in order to project the zero-year-olds in the next iteration of the projection series.

Table 2 – Count of Unacceptable Errors in Counties Between 10,000 and 30,000 Residents

County	SF1 Total Population	Unacceptable Age/Sex Bin Error Count
Aitkin County	16,202	3
Cass County	28,567	1
Chippewa County	12,441	2
Cottonwood County	11,687	5
Dodge County	20,087	3
Faribault County	14,553	4
Houston County	19,027	1
Hubbard County	20,428	2
Jackson County	10,266	5
Kanabec County	16,239	4
Koochiching County	13,311	7
Lake County	10,866	5
Martin County	20,840	1
Mille Lacs County	26,097	1
Pennington County	13,930	4
Pine County	29,750	1
Pope County	10,995	11
Redwood County	16,059	3
Renville County	15,730	1
Roseau County	15,629	2
Sibley County	15,226	3
Wadena County	13,843	1
Waseca County	19,136	1

Watonwan County	11,211	4
Yellow Medicine County	10,438	6

When we look at the larger picture, the data provided above show that 44 counties in the state will have insufficient data to produce population projections with a cohort component method. That represents over 50 percent of the counties in the state. That cannot be acceptable to the Bureau. Governmental units need these data for planning and resource allocations regardless of the size of the area in question.

Unacceptable error is not limited to counties with fewer than 30,000 residents. Even counties with significantly larger populations have some unacceptable errors, but those seem to be less pronounced, i.e. there are fewer unacceptable errors per county. In counties over 30,000 residents, 10 counties have unacceptably large errors in at least on age/sex bin. That is just shy of 29 percent of large counties in the state.

Having accurate county level data for projection is crucial as most MCDs in Minnesota, as I have pointed out to the Bureau on multiple occasions, have less than 1,000 residents. Given the issues presented here, I have no confidence in the ability of researchers and data users to be able to produce subcounty projections for anything other than our few largest cities. This will be due to the intentional perturbation of the 2020 census data.

I appreciate the Bureau taking feedback on the 2020 data products. Please let me know if you would like any additional detail on anything I have provided here. I have no issues with the Bureau sharing this feedback in any venue where it would be useful.

Sincerely,

Eric Guthrie

Eric A. Guthrie, PhD
Senior Demographer
Minnesota State Demographic Center

7. Deborah Stein, Partnership for America's Children

To whom it may concern:

I am submitting the attached comments about the implications of differential privacy on children on behalf of the Partnership for America's Children. They are based in large part on an analysis of that file by Dr. William O'Hare.

We are very concerned that the August Demonstration File continues to show significant problems about how differential privacy affects the count of children.

If you want to discuss our recommenations, please feel free to reach out to me, to my colleague Jasmine, or to Dr. O'Hare.

Sincerely,
Debbie Stein
Network Director
Partnership for America's Children

PARTNERSHIP FOR AMERICA'S CHILDREN
State and local child advocates working together for a better future

My name is Deborah Stein and I'm submitting these comments on behalf of the Partnership for America's Children, which is one of the leaders of Count All Kids.

Thank you for creating the August Demonstration File. I want to draw your attention to conducted by Dr Bill O'Hare available at our website CountAllKIds.org https://countallkids.org/new-report-analysis-of-census-bureaus-august-2022-differential-privacy-demonstration-product/ . His research found many smaller geographic areas have high levels of error in their data on young children after DP is applied. The errors are so large that they could have important implications for federal and state funding received by schools and for educational planning. Errors of this magnitude might impact formula funding that is based on Census-derived data such that some schools would get less than they should by law. It could also distort demographic predictions of school population, affecting plans for school buildings and class size. We urge the Census Bureau to try and reduce or eliminate these large errors.

We also note that the August demonstration product continues to produce highly implausible results. It shows 163,077 blocks nationwide (1.5 percent of all blocks) that had population ages 0 to 17, but no population ages 18 or over, compared to 82 such blocks before DP was applied . This unlikely large number of blocks with children and no adults may undermine confidence in the overall Census results.

These implausible results are likely due to young children being separated from their parents in 2020 Census DHC processing with differential privacy. This separation of children and parents in data processing is an ongoing concern for data on young children and the production of future tables for children. To understand the well-being of children, it is critical to understand the situation of a child's parents or caretakers.

Moreover, if the same separation of children from their parents and caregivers occurs in the application of DP to the American Community Survey, it will eliminate reliable child poverty data which is based on household income. That is, since most children have no personal income, we measure child poverty by the income of the household they live in. If the formal privacy approach applied to the American Community Survey breaks the link between children and adults in a household, we will no longer be able to measure child poverty. Child poverty rates are one of the most important measures of child well-being. Accordingly, we continue to urge that as the Bureau develops formal privacy processes for the American Community Survey and other surveys such as the CPS, it keep the data on children and adults in the household connected.

8. Jan Vink, Cornell Program on Applied Demographics

Please find attached out findings from comparing the August 2022 demonstration data with the 2010 SF1.

Jan Vink
Extension Associate
Cornell Program on Applied Demographics

Feedback on the demonstration data sets released in August 2022

Authors: Jan Vink and Leslie Reynolds
Cornell Program on Applied Demographics
September 26, 2022
Email: padinfo@cornell.edu

1 INTRODUCTION AND CONCLUSIONS

On August 25’th the Census Bureau released a new set of demonstration data with tables planned for the DHC. With only 30 days available for feedback our analyses cover a lot of topics, but due to the time constraint not all analyses could go as deep as we would help liked and although every effort is taken to present our results clearly, there was no time for finishing touches.

We didn’t attempt to compare this release with the previous release as we wanted to focus our comments on accuracy and usability of the August release.

General impressions and conclusions:

- There are some big differences in accuracy and usability between incorporated places and unincorporated places of the same size. This is very worrisome as many data uses will require custom built geographies. CDPs (unincorporated places) are examples of custom built geographies with a Census Program (PSAP) supporting this custom build. Often these custom built geographies are relatively small, less than a few thousand people and the CDP analysis show that there often big differences between the demonstration data and SF1 in those geographies.
- The Census Bureau already recognized some shortcomings in the sex/age distributions, but besides those, the age/sex distributions for the total population look usable for on/near-spine geographies (blocks groups and above) above about 1,000 persons.
- Differences in median age get smaller with growing population size.
- Differences in median age for off-spine geographies (like CDPs) are much bigger than for the on/near-spine geographies.
- Many geographies have an multi-age age range for which the differences between demonstration data and SF1 have the same sign, resulting in a significant aggregation of those differences impacting the usability.
- Breaking the link between persons and households generates many inconsistencies in the data. Census data users use various indicators that rely on the data of both households and persons:
    - Persons per Household (household population divided by number of households),
    - headship rates (householders of a certain age divided by household population of that age) are used to link age structures of the population with number of households,
    - minority home ownership (householders of a minority group divided by population of that group)

  Inconsistencies in the data require data users to derive alternative estimates that are feasible, but not longer solely based on area specific Census Data. These inconsistencies can also be seen as symptoms of distortions of the underlying distributions.

  Inconsistencies happen on all levels of geography analyzed (county and below)
- Table cells for household type and household size that generally have lower counts (less common household types and household sizes) often have very large percentage errors (e.g. over 30% of tracts have more than 10% error) which severely limits the usability of these tables.
- The householder race/ethnicity tables were mentioned in Census Bureau presentations as tables that showed lower accuracy. Our analyses confirm that and would urge the Census Bureau to improve accuracy. This problem is even more prominent under the householders of households with 3 or more generations.
- Table cells with the largest value have a negative bias causing tables to be more ‘diverse’ in the demonstration data.
- Various selection biases that are not well-understood exist within these data. For example: areas with mostly rental occupied houses have a positive bias for households with children, whereas areas with mostly owner occupied houses have a negative bias for households with children.

Analyses are based on the data downloaded directly from the Census Bureau servers or on data downloaded from NHGIS-IPUMS[1]. Several of the analyses are limited to New York State.

[1] David Van Riper, Tracy Kugler, and Jonathan Schroeder. IPUMS NHGIS, Privacy-Protected 2010 Census Demonstration Data, version 20220825 [Database]. Minneapolis, MN: IPUMS. 2022.

2 CROSSWALK

We appreciate the added tables, especially bringing the level of geography down for some often-used tables.

We would like to make two additional comments:

- Table PCT1 and H7 contain the same information and it might be confusing that table PCT1 is not available for all geographies, but table H7 is. I suggest eliminating table PCT1
- PCT13 (Age of population in Households) and its iterations are very useful, but is currently limited to iterations A-I. It would be even more useful if iterations J-O would be added (just like table P12)

3 AGE DISTRIBUTIONS

3.1 RESEARCH QUESTION:

How do differences in median age vary by type of geography, size of population, sex and race/ethnicity?

3.2 CONCLUSIONS:

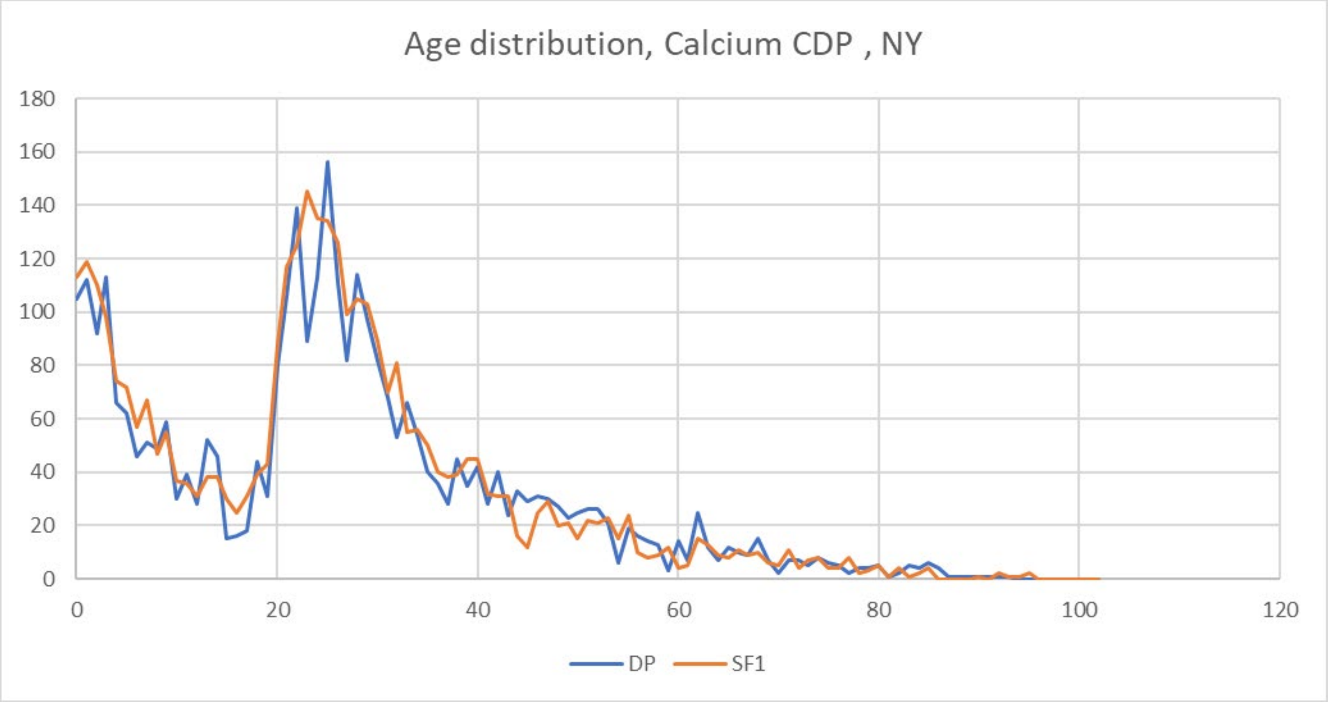
- Very small population groups (< 200) seem to have a significant negative bias
- Unincorporated places (CDPs) have larger differences compared to Incorporated places (villages and cities).
- CDPs below 2,000 population often have absolute difference of 2 year or more. This is rare among other geographies of size 500 or more
- The difference between CDPs and villages is most prominent under NH White Alone populations, not as much difference for Black Alone or Hispanic populations
- For population groups larger than 1,000 the mean absolute difference is mostly less than 1 year
- The mean absolute difference in median age for males or females where there are 1,000 to 2,000 males or females is smaller then total populations of 1,000 to 2,000
- The mean absolute difference for groups of two or more races that are 1,000 to 2,000 in size are larger than for other race categories that size

3.3 METHODS & METRICS

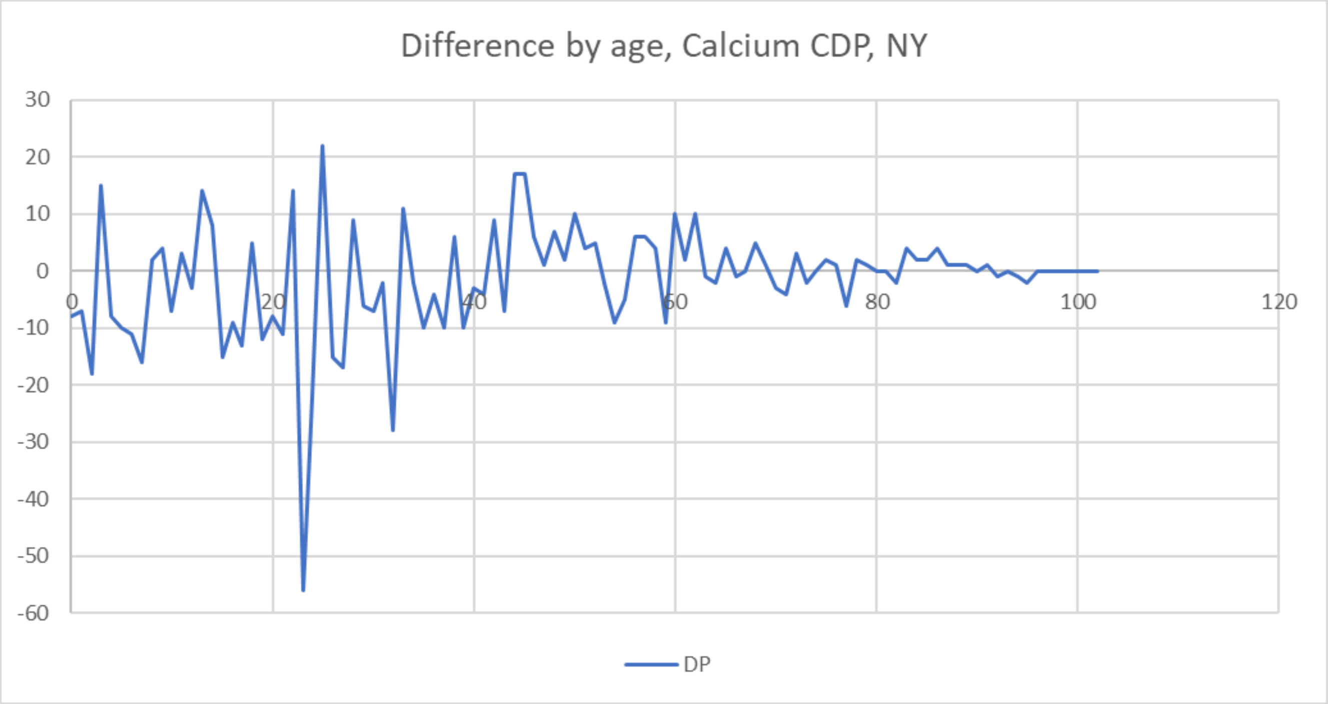
There are several ways to compare age distributions and say something about the differences.

I first use an example to explain the two methods and metrics used in this feedback. More details on the metrics are in the sections.

Example Calcium CDP:



Visually it is hard to tell if there any problems caused by DP.

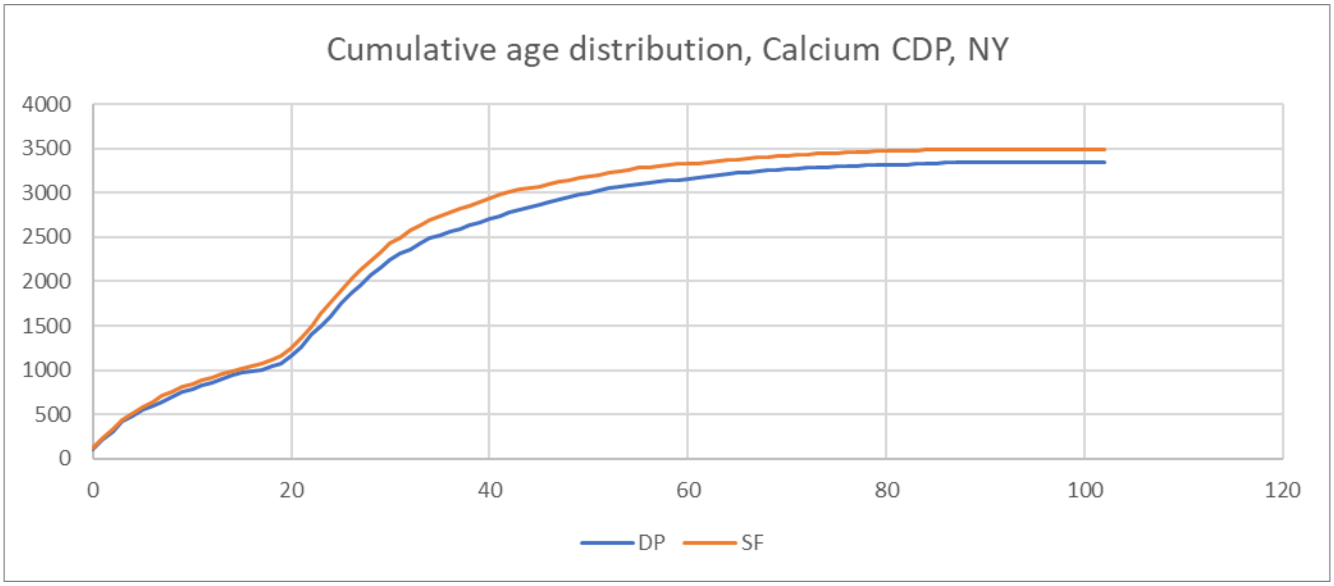


Even if we look at the differences it is hard to tell if there are any problems.

The two hypothesis I want to test are:

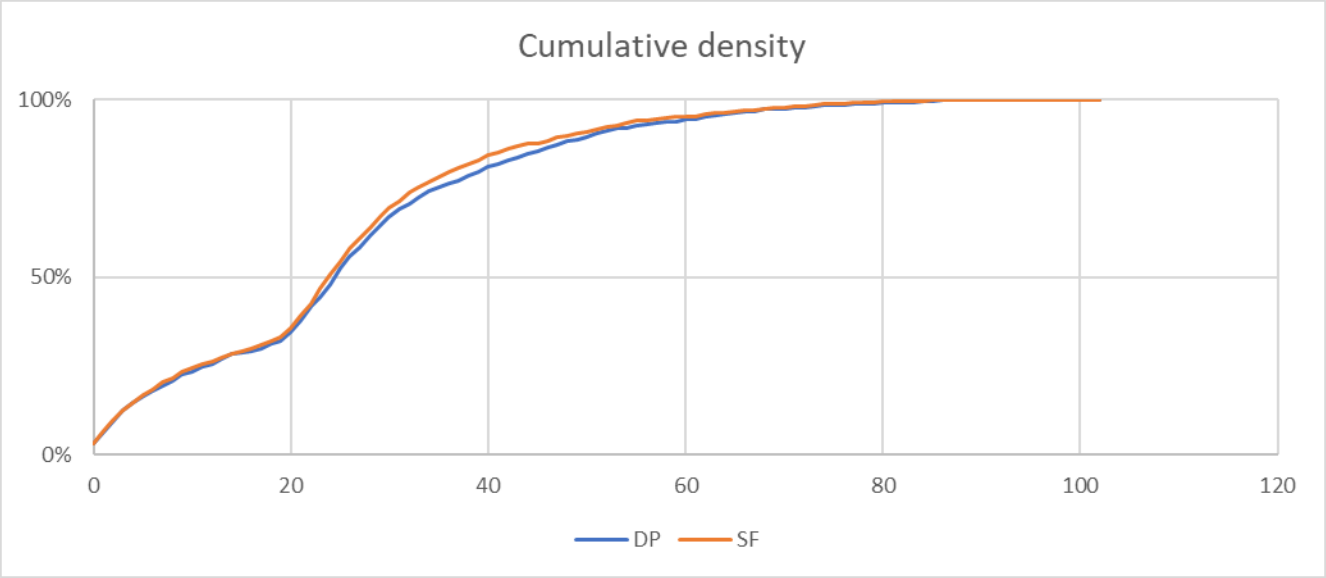
- Is there much difference in the median age?
- Is there is a group of ages for which the demonstration data yield significant different results as the published SF1 data?

To explain the methods and how they relate I start with the cumulative age distribution



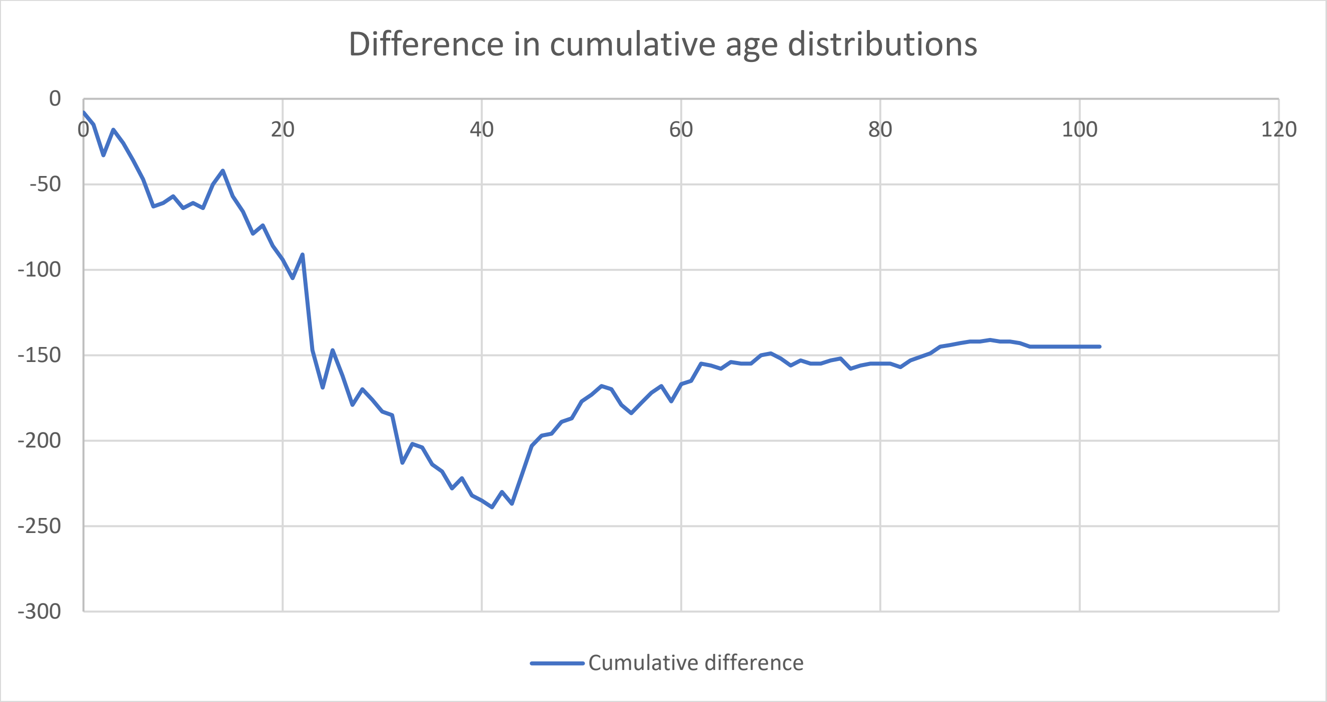
From the chart above we learn that for example in the demonstration data there were 2710 people age 0-40 and 2945 in SF1.

We can also create a cumulative density plot by dividing both distribution by the total:



From this plot we learn that the 50% of the population in Calcium is 24.5 yrs or younger in the demonstration data and 23.8 years or younger in SF1. These are the median ages and we look at the differences in median age first.

Going back to the first cumulative plot, we can also plot the difference between those two lines:



We would get the same chart if you calculate a cumulative of the errors by age.

Please note that the slope of a part of this line is mean error over that part. If we look for example between age 40 and age 60, we see a positive slope. The demonstration data counts 448 persons age 41-60 and SF1 counted 381. Indeed the positive slope corresponds with demonstration count higher then in SF1. This chart also makes it much clearer that the negative differences in the age range 0-40 far outweigh the positive differences, leading to the mentioned gap between 2710 (DP) and 2945 (SF1). Where the total population in this CDP was 4% lower in the demonstration data, for this age range the difference was 8%. In the second part of this chapter, we will be looking more into these differences.

3.4 MEDIAN AGE

3.4.1 Methodology

I only looked at geographies in New York State.

Types of geographies based on SUMLEV and first character of COUSUBCC and PLACECC considered are:

- Counties
- Incorporated places
- Unincorporated places,
- Towns & Cities (MCDs),
- Unified School Districts (SD).

Please note that cities in New York are an incorporated places (SUMLEV = 160) as well as an Subcounty (SUMLEV = 060). For these analyses they are their own category, but also included in the incorporated places.

For each of the subgroup I looked at the population size in SF1 and coded that in the following population size bins:

- 0-9 (Excluded from further analyses)
- 10-199
- 200-499
- 500-999
- 1,000-1,999
- 2,000-4,999
- 5,000-9,999
- 10,000+

3.4.2 Metrics

For each subgroup I calculated the difference in median age as the median age in the demonstration data minus the median age in the SF1.

For each subgroup based on geography, size of population, sex and race/ethnicity I calculated a Mean Difference (possibly indicating bias) and Mean Absolute Difference (measure of accuracy). I also tallied the number of times the absolute difference exceeded 1, 2 and 5 years which allows me to created percentage of all cases that exceeds those thresholds.

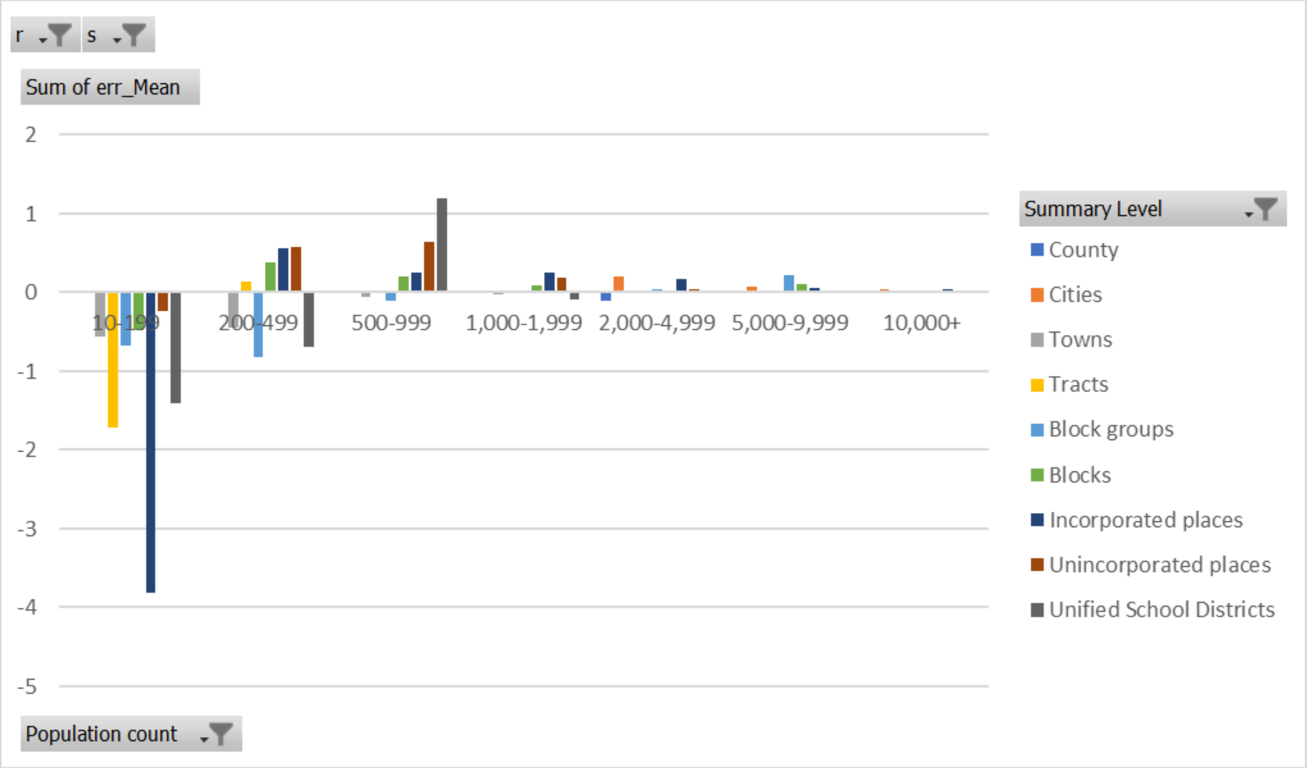
3.4.3 Results

3.4.4 Total population by geography type and size

Number of geographies by type and size

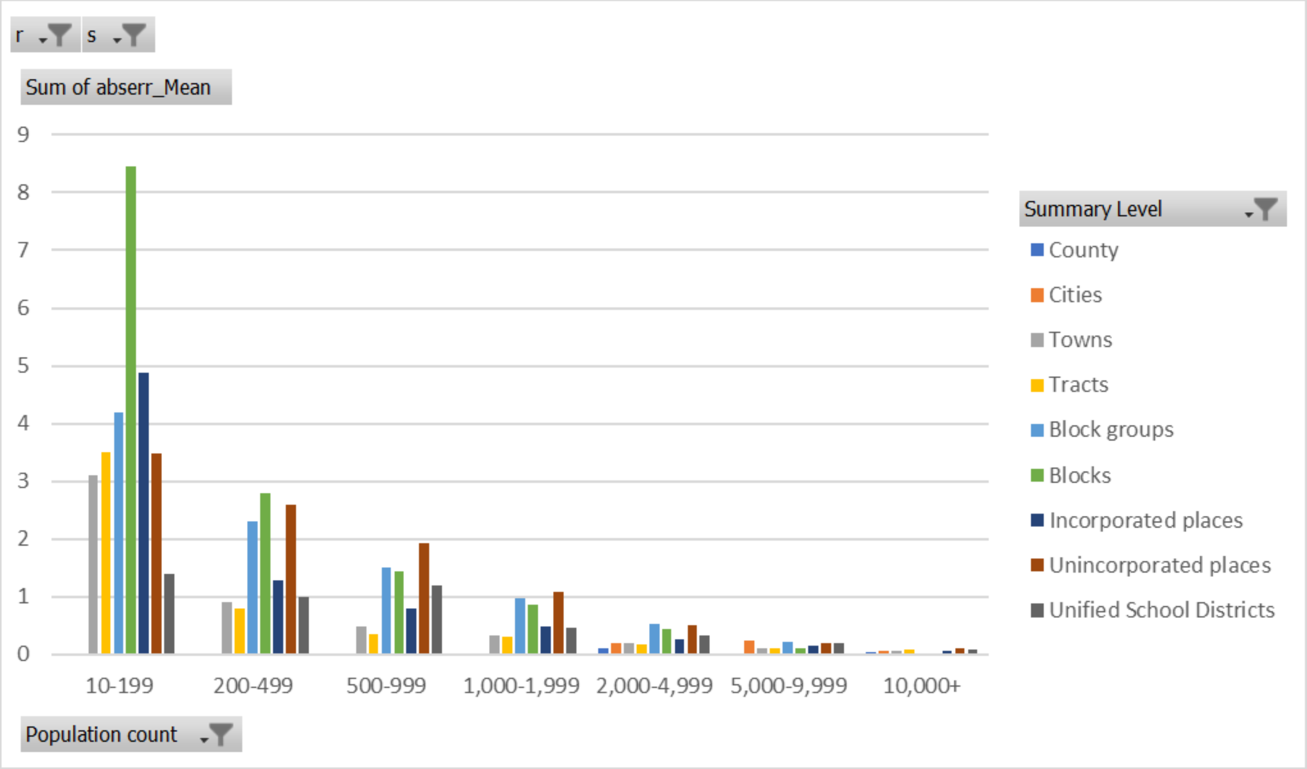
Row Labels	County	Cities	Towns	Tracts	Block groups	Blocks	Incorporated places	Unincorporated places	Unified School Districts
10-199			6	29	60	168169	6	20	1
200-499			25	12	165	14997	67	67	8
500-999			75	15	5097	4475	111	90	5
1,000-1,999			223	535	8336	1097	125	97	18
2,000-4,999	1	2	296	3000	1497	121	139	121	118
5,000-9,999		8	151	1234	14	1	82	72	198
10,000+	61	51	156	30	1		87	103	321

Figure 1: Mean difference in median age



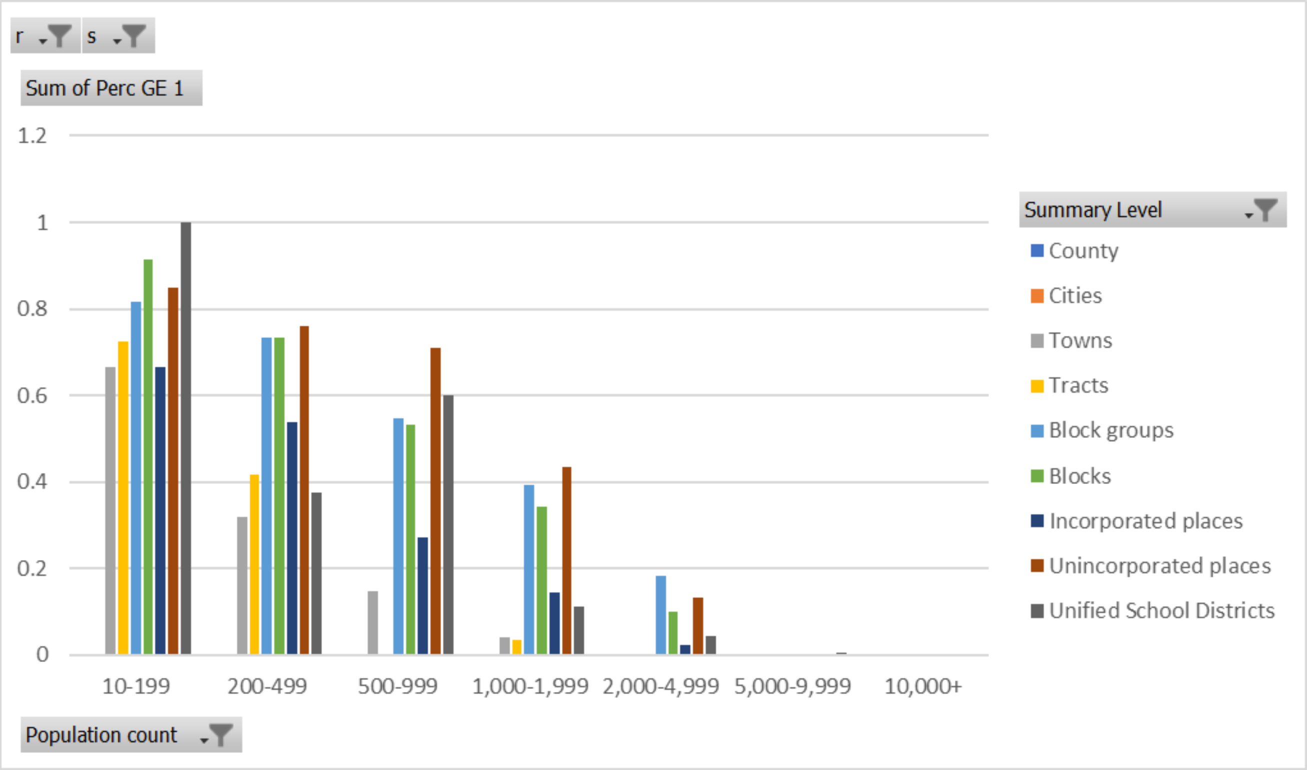
Conclusion: Very small population groups (< 200) seem to have a negative bias

Figure 2: Mean absolute difference in median age



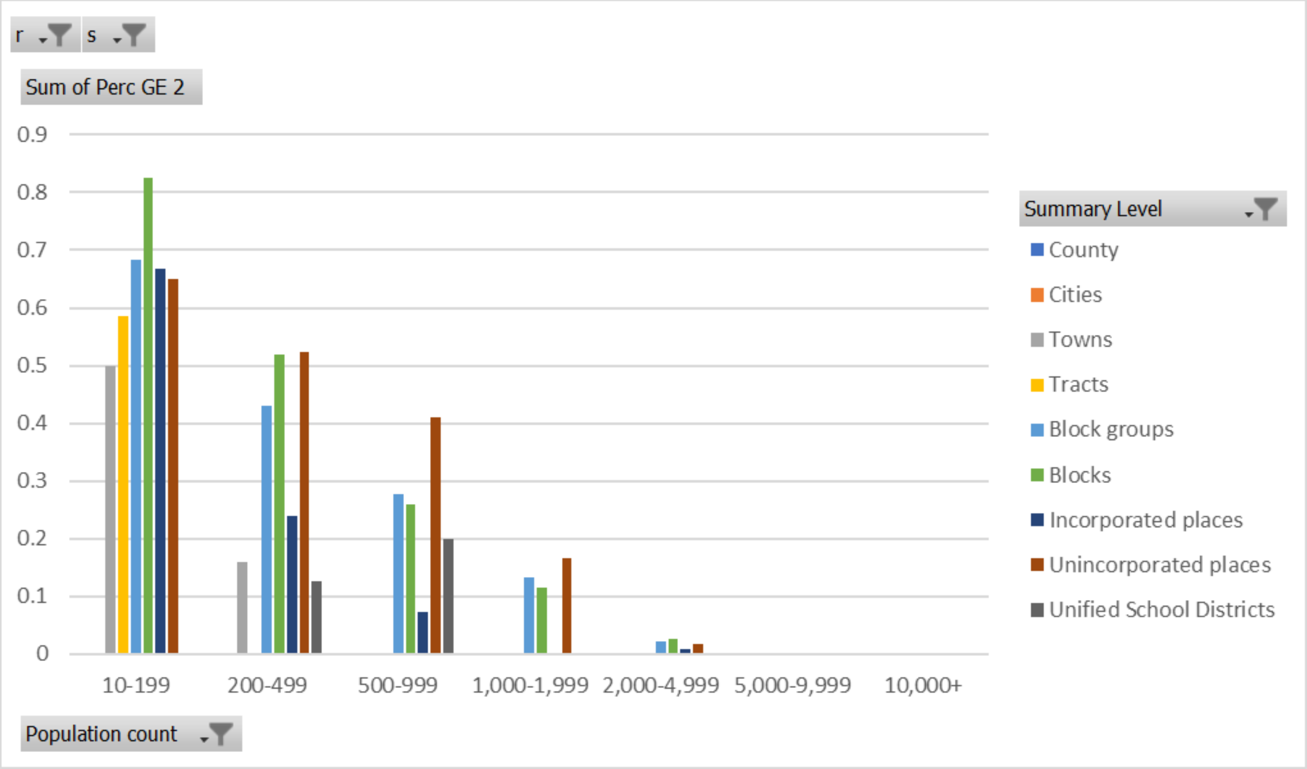
Conclusion: Unincorporated places (CDPs) have larger differences compared to Incorporated places (villages and cities).

Figure 3: Fraction of geographies with an absolute difference in median age of 1 yr or more



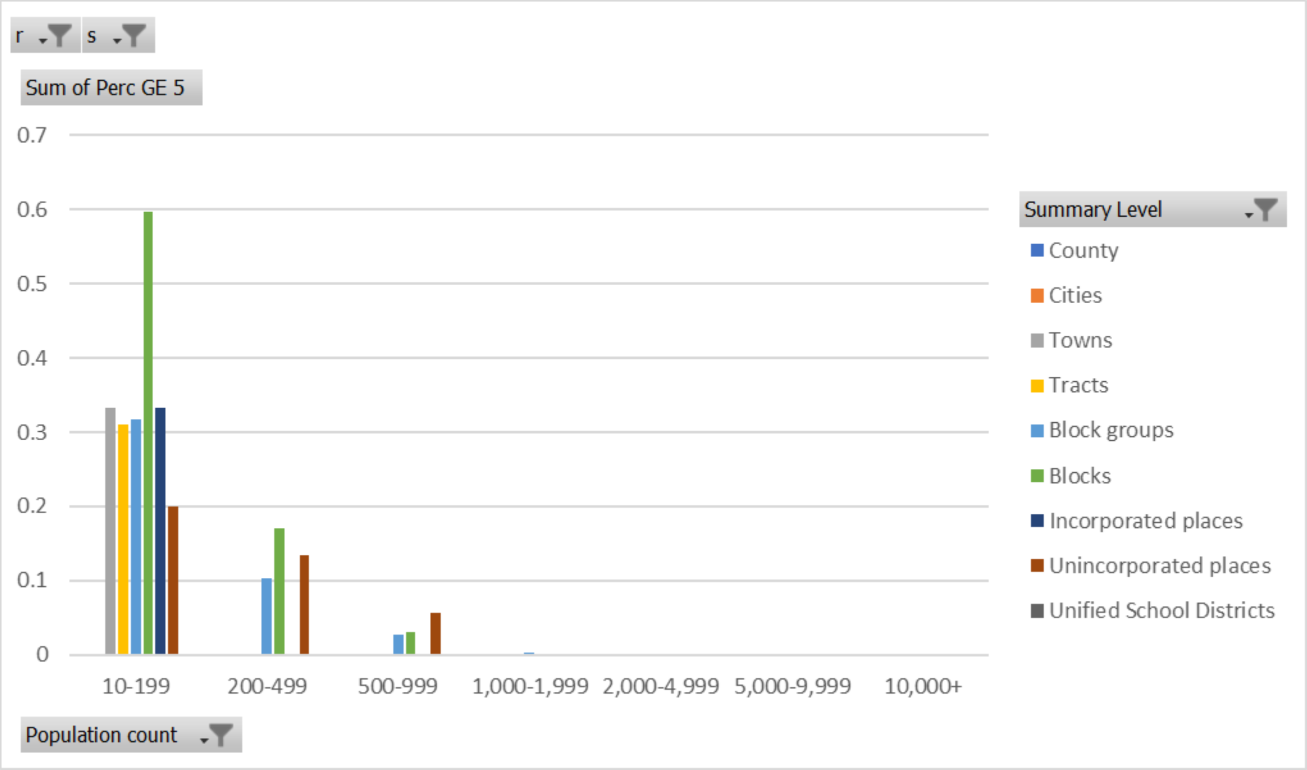
Conclusion: A 1 year difference in median age between the demonstration data and SF1 is quite common for geographies up to 2,000 persons. It is an indication that there is a gap in the cumulative age distributions and this might lead to usability problems.

Figure 4: Fraction of geographies with an absolute difference in median age of 2 yr or more



Conclusion: CDPs, block groups and blocks below 2,000 population often have absolute difference of 2 year or more. This is rare among other geographies of size 500 or more

Figure 5: Fraction of geographies with an absolute difference in median age of 5 yr or more



Conclusion: A 5 year difference between demonstration data and SF1 would certainly lead to usability problems. For geographies above 500 persons, this is rare. Caution is needed for CDP's with populations 500-999.

3.4.5 Mean absolute differences for major race/ethnicity groups

*Figure 6: Mean absolute difference in median age for **Non-Hispanic White Alone** populations*

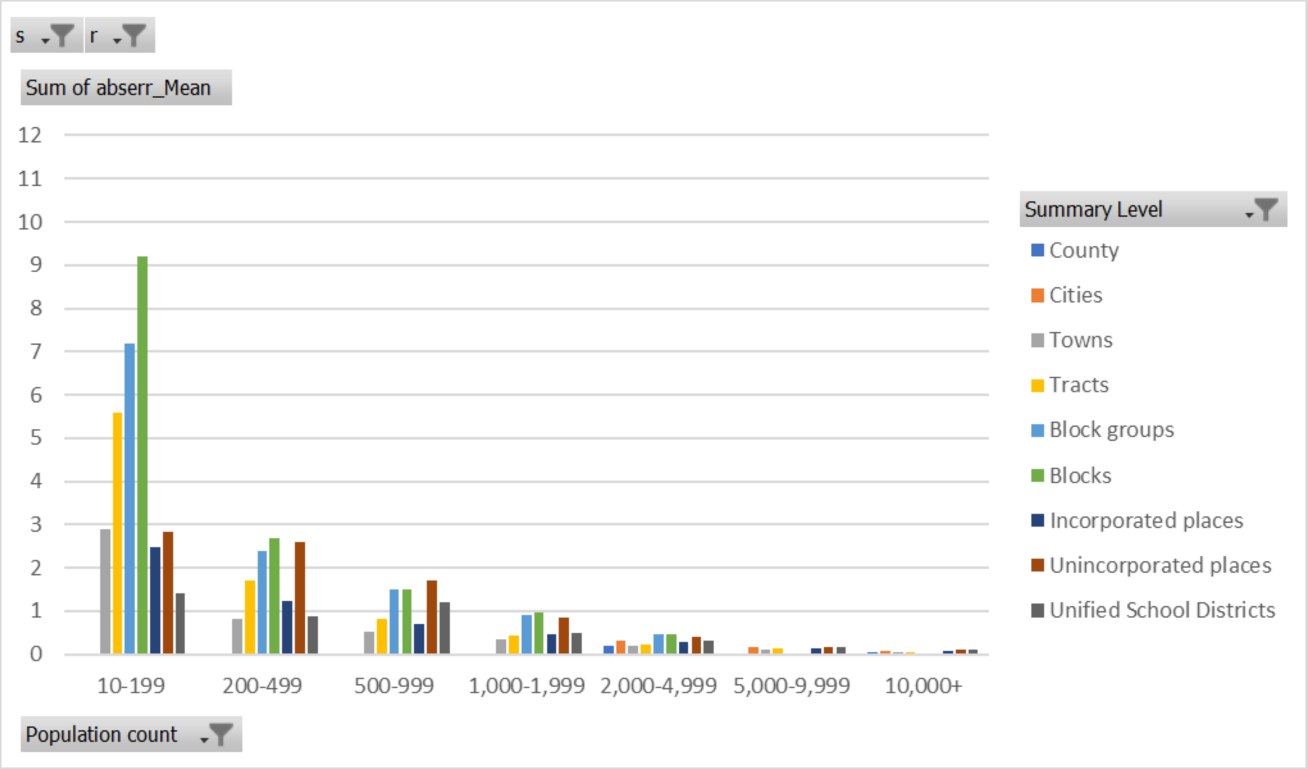


*Figure 7: Mean absolute difference in median age for **Black Alone** populations*

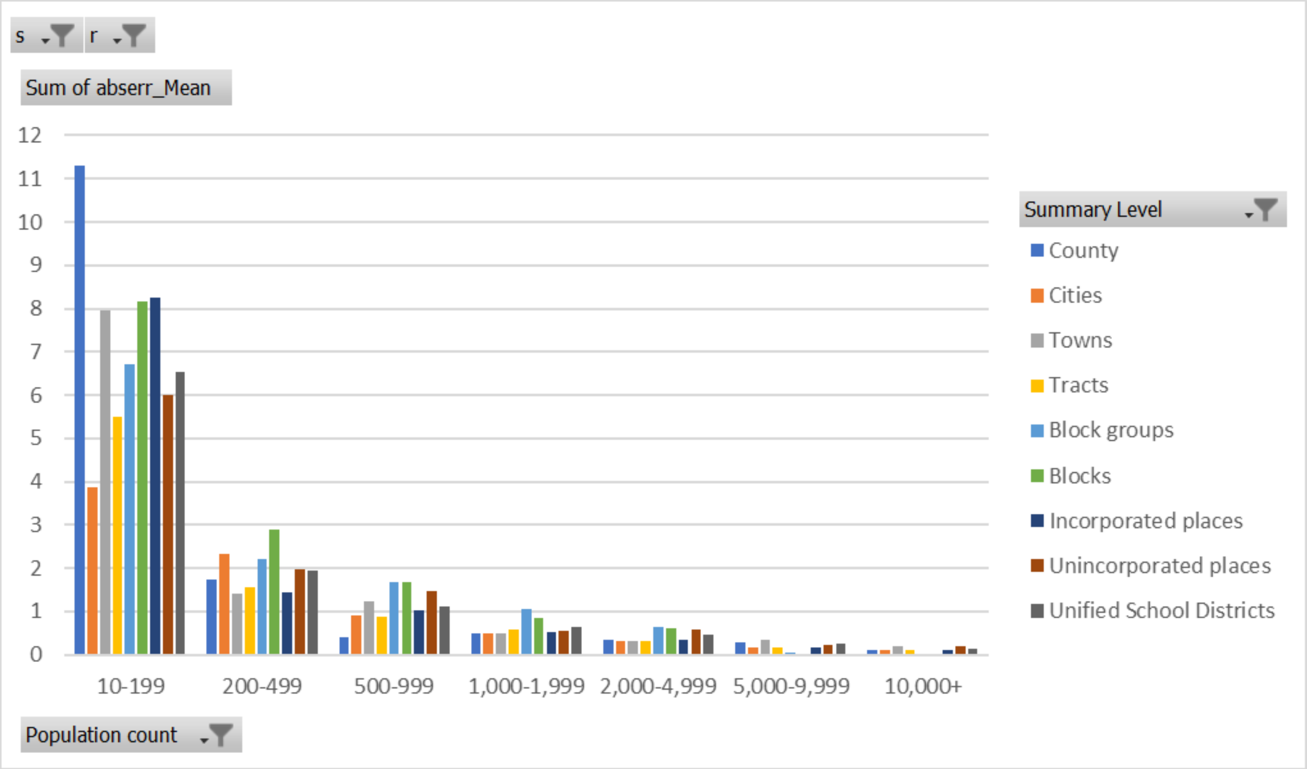
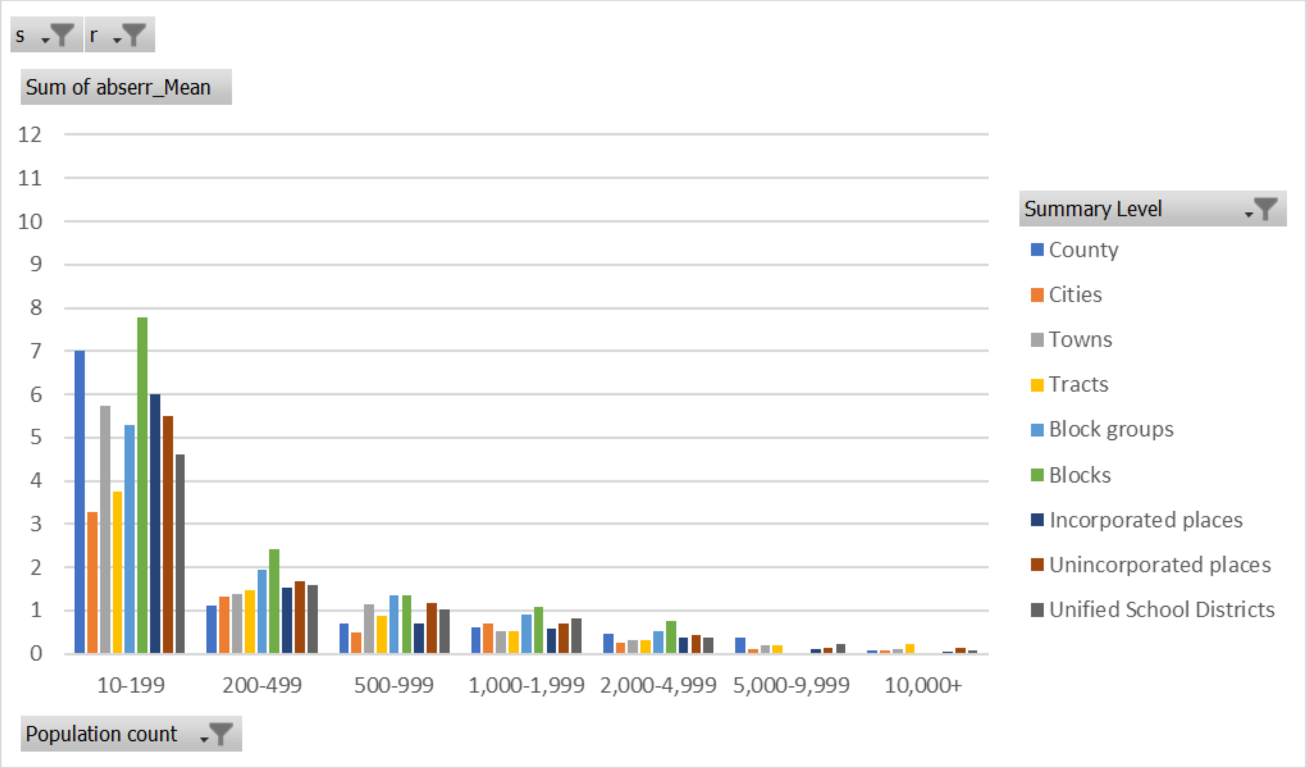


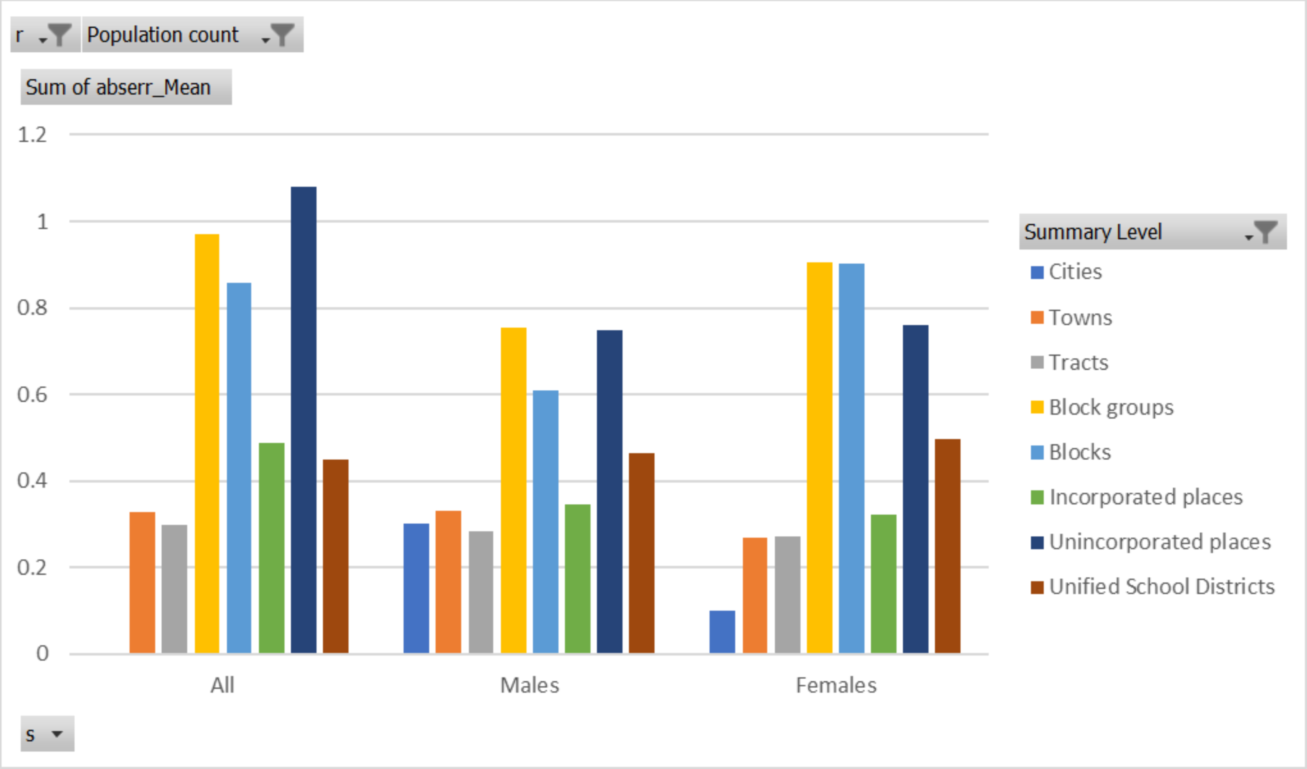
Figure 8: Mean absolute difference in median age for ***Hispanic*** *populations*



Conclusions: The differences in median age are larger for minority populations Black Alone and Hispanic than they are for Non-Hispanic White Alone populations of similar size. The differences between geographic summary levels seems to be less pronounced for Black Alone and Hispanic population.

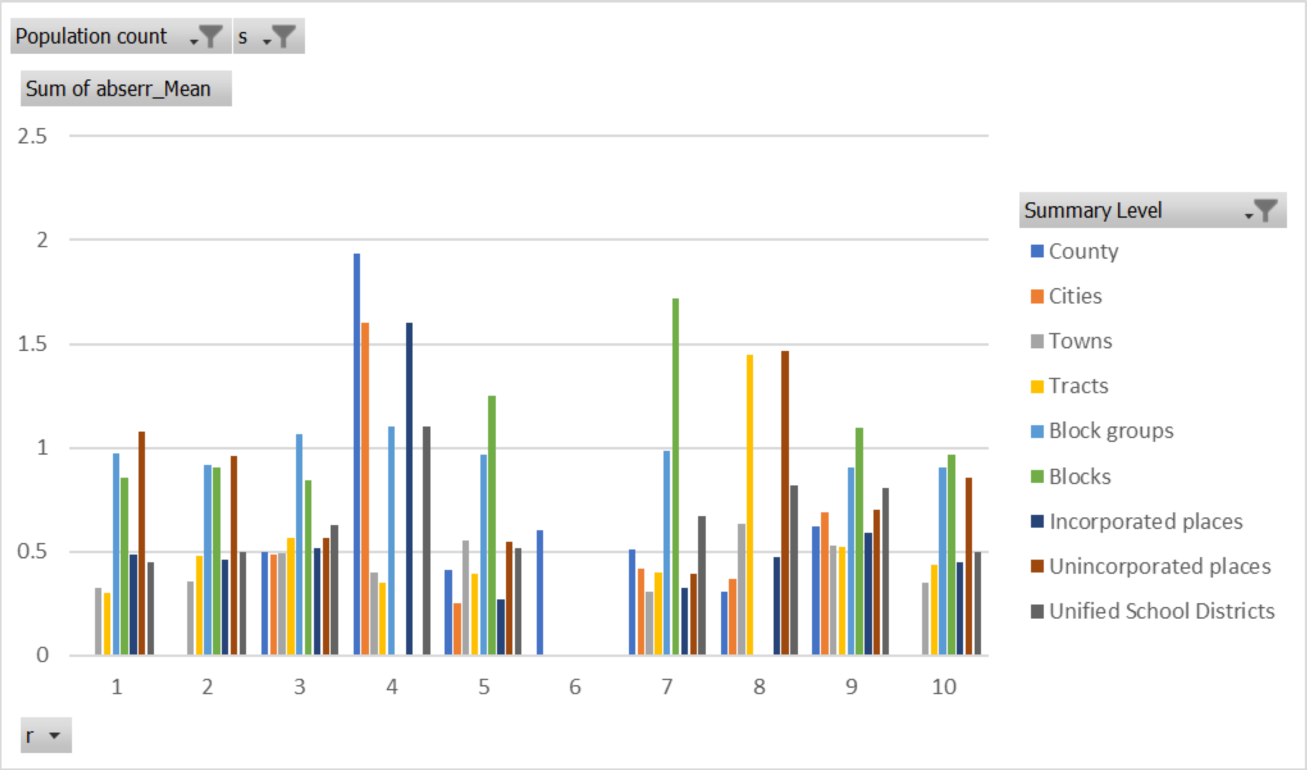
3.4.6 Mean absolute differences for population sizes 1,000 – 1,999

Figure 9: Mean absolute differences in median age by sex (group size 1,000-1,999)



Conclusion: The mean absolute difference in median age for males or females where there are 1,000 to 2,000 males or females is smaller then total populations of 1,000 to 2,000. For some geographic summary levels there is a slight difference between differences in males and females, but I suspect this difference is not significant.

Figure 10: Mean absolute differences in median age by race (group size 1,000-1,999)



Conclusion: The mean absolute difference for groups of two or more races (race code 8) that are 1,000 to 2,000 in size are larger than for other race categories that size. Black (race code 3) and Hispanic (race code 9) hve slightly higher differences than White Alone (race code 2) and NH White Alone (race code 10)

3.5 Age ranges with big differences between the demonstration data and SF1

3.5.1 Method

Data tables used: PCT12, PCT12A-PCT12O (single years of age, iterated by race categories)

Geographies selected: From the national file I selected Counties (SUMLEV = 050) and Places (SUMLEV=160). The PLACECC is used to divide the places in incorporated places (PLACECC starts with a “C”) and unincorporated places (PLACECC starts with “U”)

The observations are put in 5 bins based on SF1 universe size:

1. 500-999
2. 1,000-1,999
3. 2,000-4,999
4. 5,000-9,999
5. 10,000 plus

For each comparison, I kept tract of the cumulative difference between the demonstration data and SF1:

$$Cumulative\ difference\ at\ age\ x = \sum_{a=0}^{x} Demo(a) - SF1(a)$$

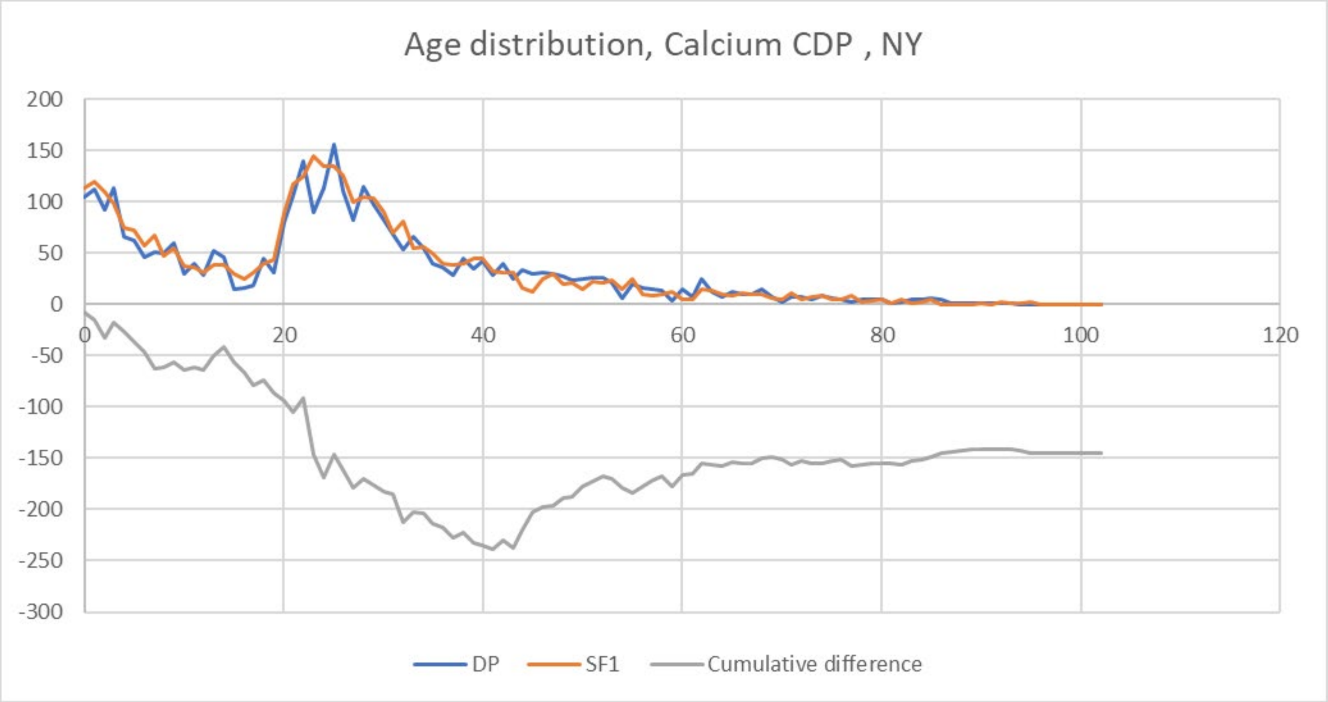
I then looked at the range of these values (max – min) as an indication of the existence of an age range where the cumulative difference went from the minimum value to the maximum value or the other way around. If the range is large than the average error within that age range is significant different from zero.

To determine what a large range is I compared it with two other metrics based on the tables:

- Comparison with the Mean Absolute Error over all ages. If the range is for example 30 times the mean error it indicates that there were at least 30 times differences with the same direction within the a limited age range. After examining the results, I found that this comparison yielded results that were indicative of statistical significance, but not necessarily meant that there would be usability issues
- Comparison with the maximum count in the table. The idea behind this is that when you plot the age distributions and the cumulative differences in a single chart, the plot of the cumulative differences should not have a much bigger range than the range of the age distribution itself.
  NOTE: this method doesn't work well for cases where the population is concentrated in a limited age range

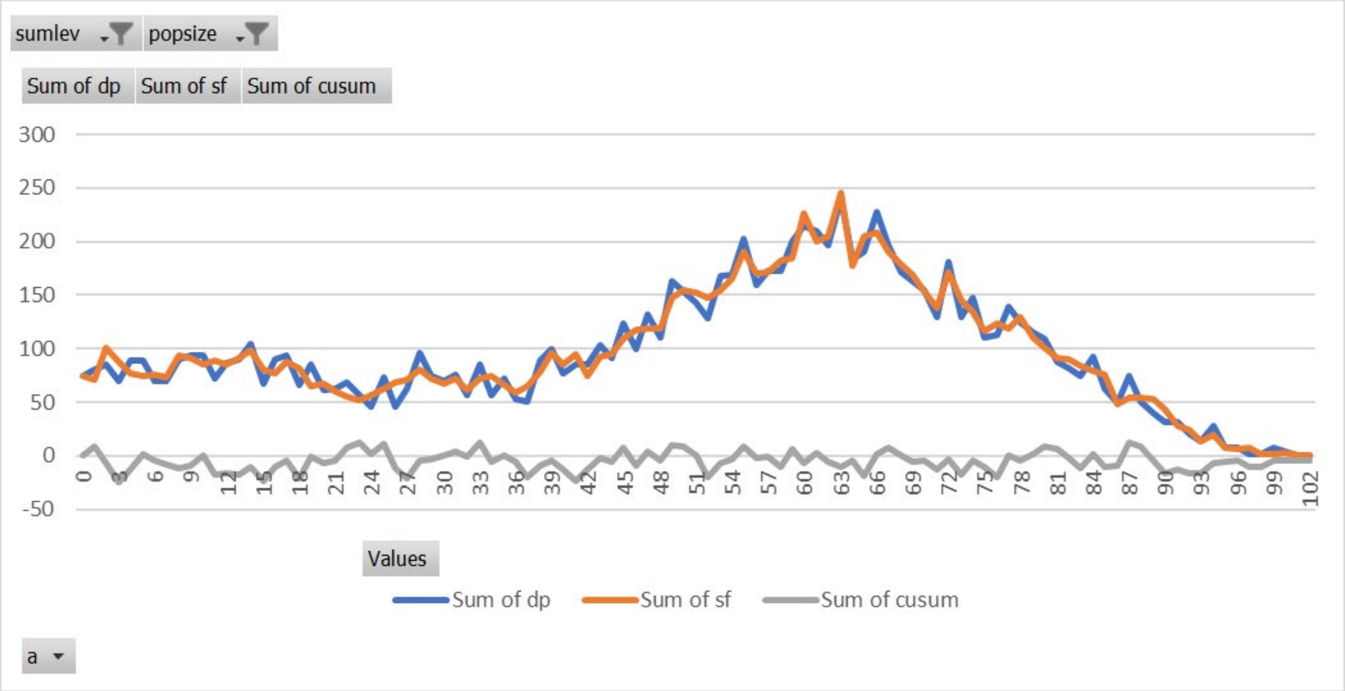
In the example of Calcium CDP I used earlier in this chapter, the cumulative differences almost reaches -250, whereas the peak of the age distribution is close to 150.

Figure 11: Age distribution in Calcium CDP, NY



Underneath is an example of the age distribution of the female population in Llano County, TX

Figure 12: Female age distribution in Llano County, TX



The range of the cumulative differences is 12-(-25)=37 and the maximum count is almost 250. In this case we don't find an age range with a lot of difference between the demonstration data and SF1.

The metric "Range cumulative difference / Max count" doesn't have an easy relation with usability, but it seems that when it exceeds 1.5 it becomes easier to identify an age range where the difference between demonstration data and SF1 could cause usability issues and the larger the value, the easier that gets.

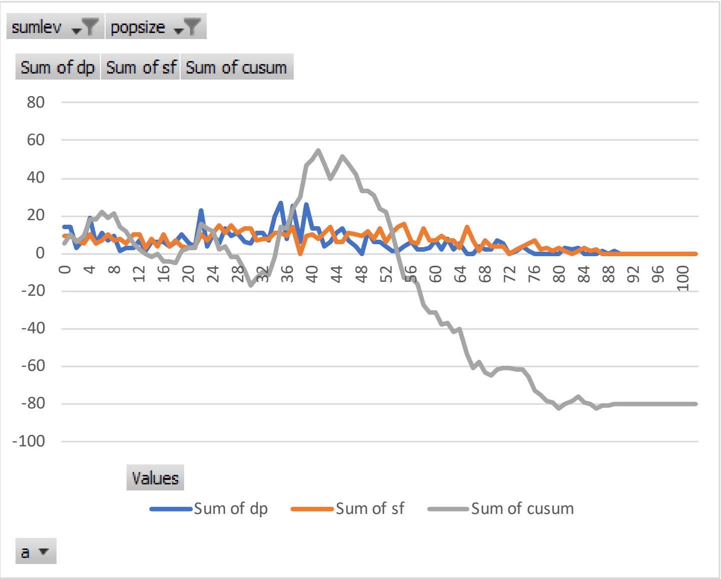
For each of the summary levels and population sizes I measured the percent of observations where the ratio exceeded a variety of thresholds.

			Ratio Range cumulative difference / Max count in SF1				
sumlev	**popsize**	**N**	**>= 1.5**	**>= 2**	**>= 2.5**	**>= 3**	**Maximum**
Counties	500-999	8902	70.2%	42.5%	21.5%	10.0%	7.64
Counties	1,000-1,999	8616	36.2%	12.2%	3.4%	1.1%	5.22
Counties	2000-4,999	11303	4.7%	0.8%	0.1%	0.0%	4.06
Counties	5,000-9,999	8875	0.1%	0.0%	0.0%	0.0%	1.87
Counties	10,000+	23232	0.0%	0.0%	0.0%	0.0%	1.37
Incorporated places	500-999	41860	33.4%	11.8%	3.7%	1.3%	7.00
Incorporated places	1,000-1,999	33571	9.4%	2.0%	0.4%	0.1%	5.25
Incorporated places	2000-4,999	31537	0.8%	0.1%	0.0%	0.0%	3.98
Incorporated places	5,000-9,999	15758	0.03%	0.0%	0.0%	0.0%	1.84
Incorporated places	10,000+	20434	0.01%	0.0%	0.0%	0.0%	1.71
Unincorporated places	500-999	19539	75.8%	51.3%	29.7%	16.7%	8.56
Unincorporated places	1,000-1,999	16307	46.9%	22.5%	9.6%	3.9%	5.81
Unincorporated places	2000-4,999	14322	13.0%	3.4%	0.7%	0.3%	4.39
Unincorporated places	5,000-9,999	5898	0.5%	0.1%	0.0%	0.0%	2.10
Unincorporated places	10,000+	4130	0.02%	0.0%	0.0%	0.0%	1.94

It is clear that geographies with larger populations have less cases where the ratio indicates possible problems. Unincorporated places have many more problems than incorporated places.

A few examples of geographies that caused the maximum ratio:

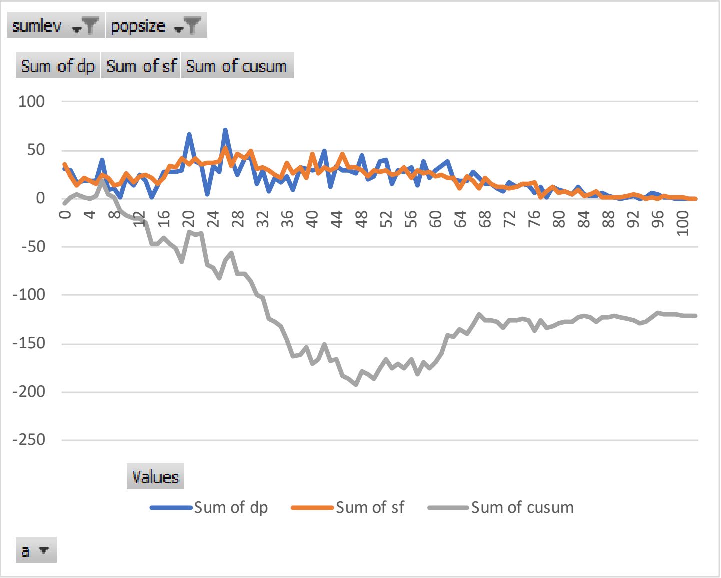
Figure 13: White Alone age distribution in Saxapahaw CDP, NC



SUMLEV		160U		
GEOID		1600000US3759580		
Name		Saxapahaw CDP		
Iteration		A		
Sex		Male		
Total pop	DP	557	Diff	-80
	SF	637		-12.6%
Median age	DP	33.6	SF1	36.7
Cumulative Error			**Ages**	**45-80**
Minimum		-82	DP	121
Maximum		55	SF1	245
Range		137	Diff	-124
Max SF1		16		-50.6%
	Ratio	8.56		

The male White Alone population in Saxapahaw CDP in North Carolina between the ages and 45 and 80 was 50% lower in the demonstration data compared to SF1

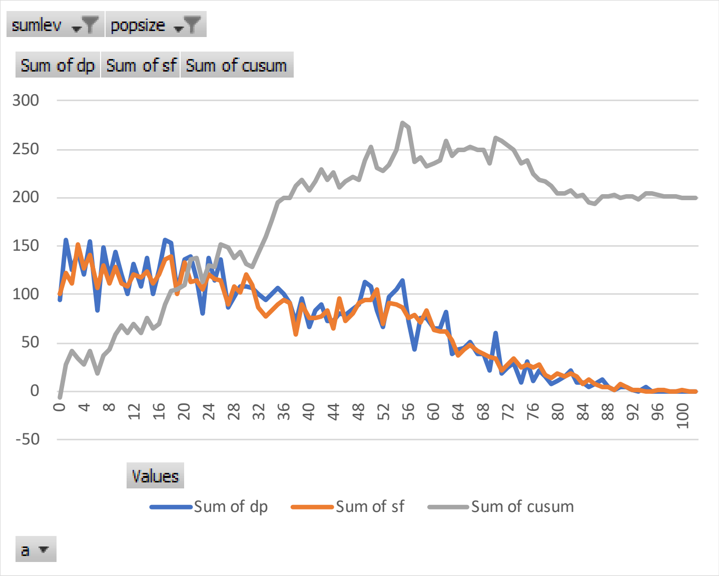
Figure 14: Non Hispanic Black age distribution in New York County, NY



SUMLEV		50		
GEOID		0500000US36061		
Name		New York County		
Iteration		K		
Sex		Total		
Total pop	DP	2023	Diff	-121
	SF	2144		-5.6%
Median age	DP	39.0	SF1	35.6
Cumulative Error			**Ages**	**5-48**
Minimum		-193	DP	1138
Maximum		18	SF1	1331
Range		211	Diff	-193
Max SF1		52		-14.5%
	Ratio	4.06		

Iteration K is Non-Hispanic Black. The total in New York County, NY [Manhattan] is 5.6% lower in the demonstration data, but between the ages 5 and 48 the difference is 14.5%.

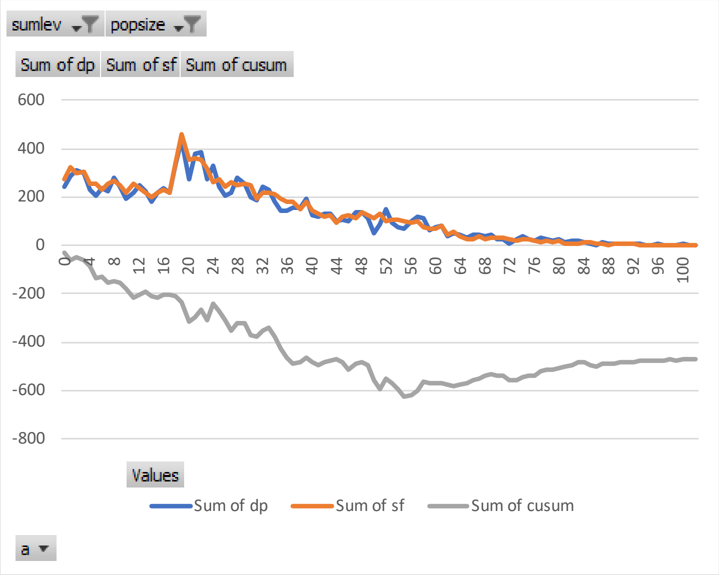
Figure 15: Non-Hispanic Black age distribution in Humboldt County, CA



SUMLEV		50		
GEOID		0500000US06023		
Name		Humboldt County		
Iteration		K		
Sex		Total		
Total pop	DP	7161	Diff	200
	SF	6961		2.9%
Median age	DP	28.0	SF1	28.4
Cumulative Error			**Ages**	**0-55**
Minimum		-6	DP	5955
Maximum		277	SF1	5706
Range		283	Diff	249
Max SF1		151		4.4%
	Ratio	1.87		

The metric for the total Non-Hispanic Black population in Humboldt County, CA is 1.87. The median age is very similar between the two data sources. For the ages 0-55 the difference is much bigger than the overall difference (4.4% vs 2.9%), but that might still be usable.

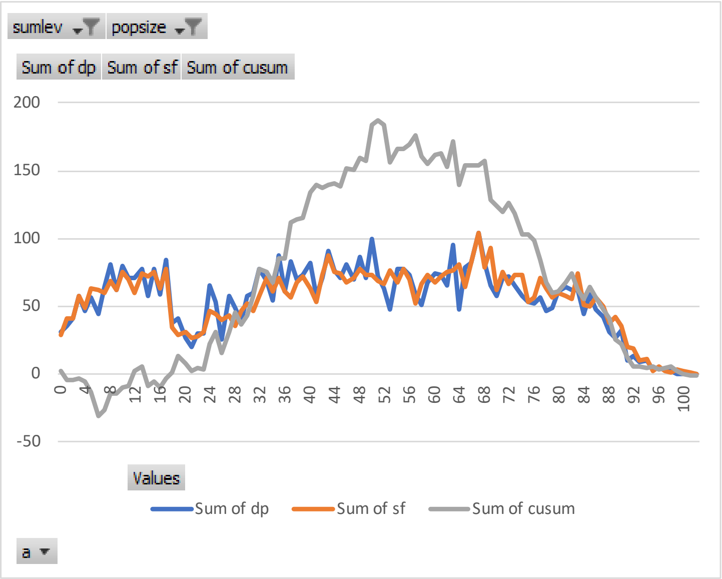
Figure 16: Hispanic age distribution in Humboldt County, CA



SUMLEV		50		
GEOID		0500000US06023		
Name		Humboldt County		
Iteration		H		
Sex		Total		
Total pop	DP	12742	Diff	-469
	SF	13211		-3.6%
Median age	DP	23.0	SF1	22.7
Cumulative Error			**Ages**	**0-55**
Minimum		-625	DP	11383
Maximum		-29	SF1	11977
Range		625	Diff	-594
Max SF1		457		-5.0%
	Ratio	1.37		

Humboldt County, CA also had the largest metric value for the over 10,000 category, this time for iteration H (Hispanic population). The difference in the 0-55 yr old population was -5%, which could cause usability problems.

Figure 17: Female age distribution in Viera East CDP, FL



SUMLEV	160U					
GEOID	1600000US1274368					
Name	Viera East CDP					
Iteration	Total					
Sex	Female					
Total pop	DP	5640		Diff		-1
	SF	5641				0.0%
Median age	DP	46.4	SF1		48.4	
Cumulative Error				**Ages**		**5-50**
Minimum		-31		DP		2810
Maximum		187		SF1		2647
Range		218		Diff		163
Max SF1		104				6.2%
	Ratio	2.10				

The age distribution female population in Viera East CDP, FL has the largest metric value for the 5,000-9,999 population size. Between age 5-50 there was a difference of 6.2% whereas the total population was right on.

4 GEOGRAPHIES WITH IMPOSSIBLE STATISTICS

4.1 RESEARCH QUESTION:

Breaking the connection between households and persons lead to many impossible statistics. How often do they appear for different levels of geography?

4.2 CONCLUSIONS:

- Block level data is full of inconsistencies
- There are a few fields that by definition should be the same in the person file as in the household file (e.g. householders = households, householders living alone = single person households). In this data set, this is very rarely the case. Big differences also occur rather often in sub-county geographies
- The number of older age householders often exceeds the population in that age group
- The number of minority householders often exceeds the population in those race groups. At the block group level the householders often outnumber the population by more than 10 and more than 10%

4.3 METHOD:

The SUMLEV field was used to determine the geographic level. Records with zero population were left out of the analyses. Several inconsistencies were flagged and if the difference exceeded 10 and 10% were flagged as a big error. The percent error is calculated as a difference divided by the average of both observations. Extreme examples were chosen by looking at a combination of the percent error and the size of the population

4.4 MORE HOUSEHOLDS THAN HOUSEHOLD POPULATION

The household population (from table H8) should be larger than the number of occupied houses (from table H3).

		Flagged		Big error	
Summary level	N	count	%	count	%
County	62	0		0	
Tract	4870	1	0.02%	0	
Block group	15194	2	0.01%	0	
Blocks	244281	4646	1.9%	84	0.03%
MCD	1010	0		0	
Place	1189	1	0.08%	0	
Unified SD	669	0		0	

Extreme examples:

Block 361019613001000: Total population = 299, Household population = 1, occupied houses = 15

Block 360550094002030: Total population = 140, Household population = 140, occupied houses = 156

4.5 HOUSEHOLD POPULATION WITHOUT OCCUPIED HOUSES

If there is household population (from table H8) than the number of occupied houses (from table H3) should be non-zero.

Summary level	N	Flagged count	%
County	62	0	
Tract	4870	10	0.2%
Block group	15194	12	0.08%
Blocks	244281	16930	6.9%
MCD	1010	2	0.2%
Place	1189	0	
Unified SD	669	0	

Extreme examples:

Block 361031456033001: Total population = 71, Household population = 71, occupied houses = 0 (out of 15 total housing units)

4.6 HOUSEHOLDERS NOT EQUAL TO HOUSEHOLDS

The population with relationship "householder" (from table P17) should be equal to the number of occupied houses (from table H3).

This is especially important when calculating Persons per Household (PPH) where we often have two different numbers for the denominator.

Summary level	N	Flagged count	%	Big error count	%
County	62	60	96.8%	0	
Tract	4870	4555	93.5%	14	0.3%
Block group	15194	14826	97.6%	1256	8.3%
Blocks	244281	223729	91.6%	47857	19.6%
MCD	1010	964	95.4%	0	
Place	1189	1150	96.7%	86	7.2%
Unified SD	669	655	97.9%	4	0.6%

Extreme examples:

Pleasant Valley CDP: Household population = 1,154,
Householders = 476, Persons per household based on householders = 2.42
Occupied houses = 541, PPH based on occupied houses = 2.13

Blockgroup 3608111551023: Household population = 1,322,
Householders = 529, PPH based on householders = 2.50
Occupied houses = 411, PPH based on occupied houses = 3.22

4.7 HOUSEHOLDERS LIVING ALONE FROM THE PERSON FILE NOT EQUAL TO HOUSEHOLDERS LIVING ALONE FROM THE UNIT FILE

The population with relationship “householder living alone” (from table P17) should be equal to the number of households with household type “Householder living alone” (from table P16).

		Flagged		Big error	
Summary level	N	count	%	count	%
County	62	62	100.0%	0	
Tract	4870	4644	95.4%	101	2.1%
Block group	15194	14842	97.7%	7346	48.3%
Blocks	192337	184599	75.6%	34141	14.0%
MCD	1010	956	94.7%	10	1.0%
Place	1189	1139	95.8%	160	13.5%
Unified SD	669	656	98.1%	12	1.8%

Extreme examples:

Crugers CDP:	Total population = 1,565, Householders living alone 243 male + 352 female = 595, Household type “householder living alone” = 521
Blockgroup 360610191001:	Total population = 1476, Householders living alone 147 male + 234 female = 381, Household type “householder living alone” = 518
Block 360470944013003:	Total population = 405, Householders living alone 2 male + 0 female = 2, Household type “householder living alone” = 56

4.8 HOUSEHOLD POPULATION UNDER 18 LESS THEN NUMBER OF HOUSEHOLDS WITH CHILDREN UNDER 18

The household population under age 18 (from table P15) should be at least as large as the number of households with one or more people under 18 (from table P21).

		Flagged		Big error	
Summary level	N	count	%	count	%
County	62	0		0	
Tract	4870	14	0.3%	5	0.1%
Block group	15194	196	1.3%	83	0.5%
Blocks	192337	45383	18.6%	3079	1.3%
MCD	1010	3	0.3%	1	0.1%
Place	1189	13	1.1%	3	0.3%
Unified SD	669	2	0.3%	1	0.1%

Extreme examples:

Block group 361190004013:	Total population = 670, Household population under 18 = 52, Households with children under 18 = 137

4.9 NOT ENOUGH HOUSEHOLD POPULATION TO FILL THE HOUSEHOLD BY SIZE STATISTICS

One can calculate an under bound for the household population from table H9 (households by size) by multiplying each size category by the size and multiply the 7-or-more category by 7. The household population (table H8) should be larger than this under bound.

One can also calculate what the average household size of the 7 or more category should be to match the household population. Values much larger than 10 are very improbable. These analyses are not part of this feedback.

Summary level	N	Flagged count	%	Big error count	%
County	62	29	46.8%	0	
Tract	4870	1777	36.5%	24	0.5%
Block group	15194	6840	45.0%	847	5.6%
Blocks	192337	102403	41.9%	49766	20.4%
MCD	1010	542	53.7%	16	1.6%
Place	1189	685	57.6%	115	9.7%
Unified SD	669	333	49.8%	7	1.0%

Extreme examples:

Block 361032010014012:	Household population = 8, under bound based on housing size = 83 (1*2+1*3+1*4+5*5+7*7) – 15 households in total which is greater than the household population
Delaware County	Household population = 45,556, under bound based on housing size = 45,843 (which is reached when all 190 7+ households have exactly 7 household members)
Thousand Island Park CDP:	Household population = 49, under bound based on housing size = 145 (11*1 + 17*2+14*3+4*4+7*5+1*7) – 54 households in total which is greater than the household population
Blockgroup 361190004013:	Household population = 670, under bound based on housing size = 1054 (138*1 + 83*2+50*3+38*4+34*5+23*6+20*7)

4.10 More householders of a certain age group than population of that age group

The number of people in an age group (from table P12) should greater or be equal to the number of householders in that age group (from table H13)

Geographies without population and without householders in a certain age group are excluded from these analyses.

Figure 18: Share of geographies with number of householders exceeding population by age group

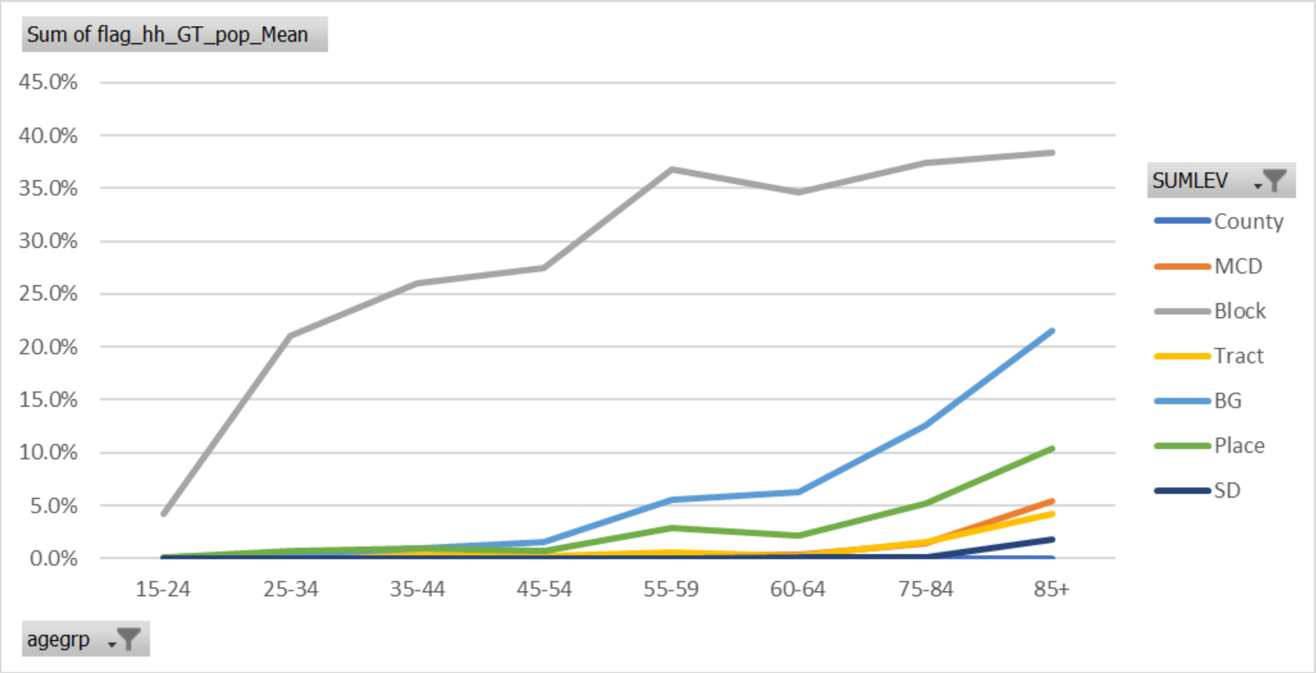
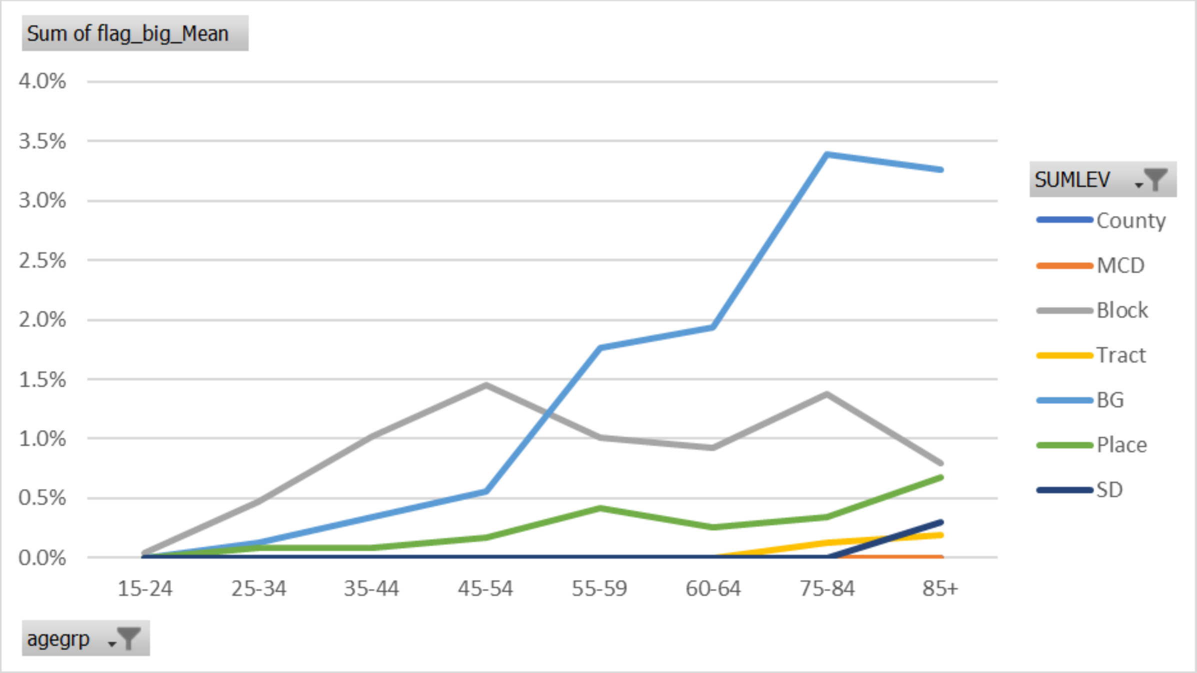


Figure 19: Share of geographies with number of householders greatly exceeding population by age group



Extreme examples:

Blockgroup 360610093005: householders age 60-64 = 53, population age 60-64 = 1

4.11 MORE HOUSEHOLDERS OF A CERTAIN RACE/ETHNICITY GROUP THAN POPULATION OF THAT SAME GROUP

The number of people in a race group (from table P5) should greater or be equal to the number of householders in that race group (from table H7)

Geographies without population and without householders in a certain race/ethnicity group are excluded from these analyses.

Figure 20: Share of geographies with number of householders exceeding population by race/ethnicity group

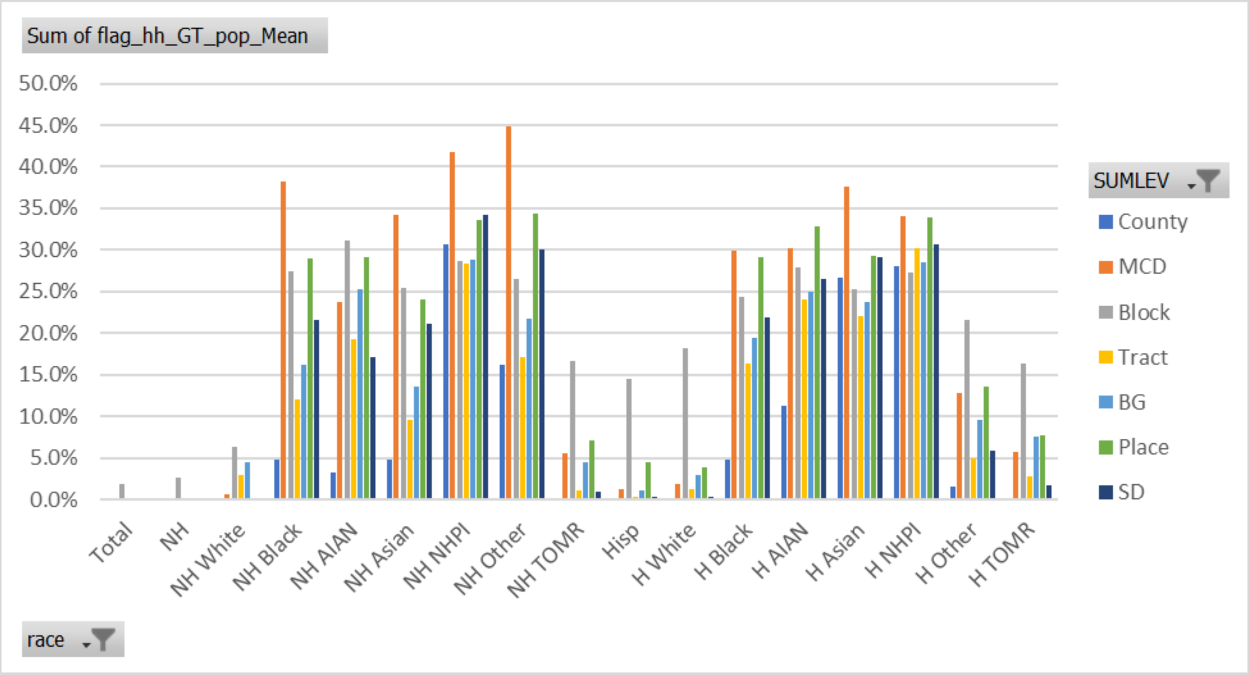
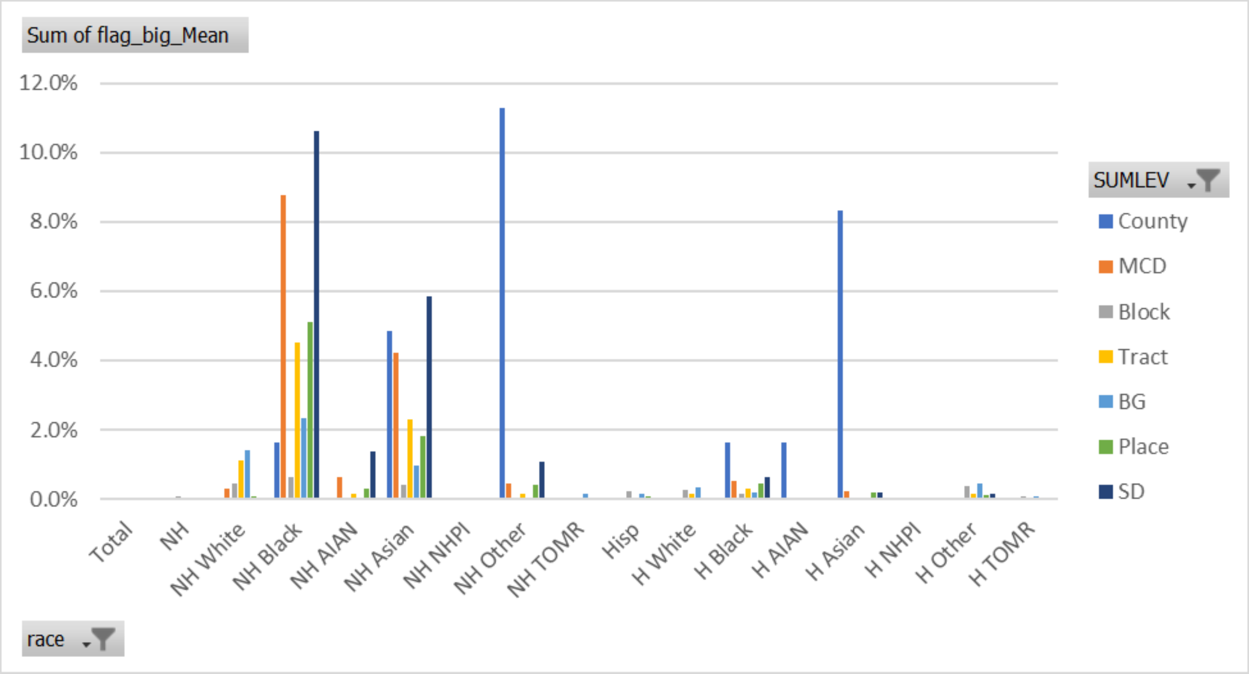


Figure 21: Share of geographies with number of householders greatly exceeding population by race/ethnicity group



Extreme examples:

Blockgroup 360610151002: NH Black Alone householders = 123, NH Black Alone population = 21

5 COMPARING INCORPORATED PLACES, UNINCORPORATED PLACES AND URBAN AREAS

Unincorporated places and urban areas are not on the traditional spine of the Top-Down Algorithm, whereas incorporated places are brought closer to the spine by creating and optimized block group geography.

In this chapter results from incorporated places are compared with unincorporated places and urban areas.

5.1 RESEARCH QUESTION:

Is there much difference in error metrics between incorporated places, unincorporated places and urban areas.

5.2 CONCLUSIONS:

- Severe errors are much more frequent for unincorporated places than for incorporated places for similar size. Urban areas look slightly better than unincorporated places, probably benefitting from the accuracy of the incorporated place at the heart of the urban area.
- Big errors are too common for household type: "Male Householder, no spouse present" and "householder not living alone". Even in incorporated places with more than 5,000 residents, differences of more than 10% were quite common.
- Larger places (regardless of type of place) showed many severe errors for the number of larger households.

5.3 METHOD:

From the national Summary File (Demonstration data and SF1 – Urban/Rural Update) I extracted geographies with SUMLEV = 160 (places) and SUMLEV = 400. For the places I used the first character of PLACECC variable to split the places in Incorporated places (PLACECC starts with "C") and Unincorporated places (starts with "U"). All places with other place types were ignored, like military places.

I further restricted my observations to those where at least 80% of the population resided in households.

The observations are put in 5 bins based on SF1 universe size:

1. 500-999
2. 1,000-1,999
3. 2,000-4,999 (Urban areas are by definition larger then 2,500)
4. 5,000-9,999
5. 10,000 plus

I chose three tables for further analyses.

- P12: Population by age and sex. I created 5 year age groups and only looked at the males+females
- P16: Household type
- H13: Household size

For each field I calculated an absolute percent error as the absolute difference between the demonstration data divided by the average of demonstration data and SF1: |DP – SF1| / 0.5 * (DP + SF1). If this absolute percent error was more or equal to 10% and the absolute difference was also more or equal to 10 then I flagged it as a big error.

The tables in this section reflect the share of the observations that showed big errors. The color is a visual aid, the darker the higher the share. All tables use the same coloring scheme with the darkest red reflecting 50% or more observations with big errors.

5.4 Results

Table 1: Share of observations with big errors in table P12: Population by age and sex

	Incorporated places						Unincorporated places						Urban areas		
	0-499	500-999	1,000-1,999	2,000-4,999	5,000-9,999	10,000+	0-499	500-999	1,000-1,999	2,000-4,999	5,000-9,999	10,000+	2,500-4,999	5,000-9,999	10,000+
Total	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	1.2%	0.3%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
0-4	0.3%	1.3%	2.3%	0.4%	0.0%	0.0%	8.1%	26.1%	29.9%	18.8%	3.0%	0.5%	7.7%	2.1%	0.1%
5-9	0.1%	0.9%	1.3%	0.1%	0.0%	0.0%	8.4%	27.2%	30.8%	15.7%	2.3%	0.4%	7.6%	1.4%	0.1%
10-14	0.2%	1.1%	1.8%	0.2%	0.0%	0.0%	10.0%	27.3%	29.4%	15.0%	2.2%	0.3%	8.6%	1.9%	0.1%
15-19	0.8%	3.2%	5.7%	1.1%	0.0%	0.0%	11.0%	32.1%	36.2%	18.2%	3.2%	0.8%	10.5%	1.9%	0.0%
20-24	0.1%	0.5%	1.6%	0.5%	0.0%	0.0%	8.2%	27.9%	33.3%	23.9%	5.4%	0.7%	11.1%	3.0%	0.1%
25-29	0.1%	0.8%	1.3%	0.4%	0.1%	0.0%	8.9%	28.4%	33.2%	23.3%	6.4%	0.7%	9.0%	2.6%	0.1%
30-34	0.1%	0.6%	1.4%	0.3%	0.0%	0.0%	9.6%	27.3%	34.1%	23.1%	5.5%	0.7%	10.4%	2.8%	0.2%
35-39	0.3%	1.4%	2.0%	0.2%	0.0%	0.0%	11.4%	33.1%	38.1%	22.5%	5.4%	0.8%	11.6%	3.2%	0.1%
40-44	0.3%	1.7%	2.3%	0.5%	0.0%	0.0%	13.8%	32.6%	40.0%	21.9%	3.5%	0.8%	12.2%	2.5%	0.2%
45-49	0.3%	1.2%	1.6%	0.1%	0.0%	0.0%	15.1%	38.4%	41.0%	18.2%	3.9%	0.1%	10.1%	1.7%	0.3%
50-54	0.3%	1.3%	1.6%	0.1%	0.0%	0.0%	17.1%	40.4%	39.2%	18.9%	3.4%	0.1%	10.0%	2.1%	0.3%
55-59	0.2%	0.9%	1.6%	0.2%	0.0%	0.0%	16.2%	39.7%	40.3%	21.4%	4.1%	0.1%	11.3%	4.6%	0.2%
60-64	0.5%	2.4%	4.0%	0.5%	0.0%	0.0%	15.7%	36.4%	39.0%	24.3%	5.9%	0.7%	13.2%	5.4%	0.4%
65-69	0.3%	1.7%	3.4%	1.7%	0.1%	0.0%	11.0%	29.2%	35.9%	27.5%	6.9%	0.8%	17.1%	6.8%	0.9%
70-74	0.0%	0.3%	0.6%	0.8%	0.1%	0.0%	7.4%	21.1%	31.4%	28.4%	10.0%	1.9%	16.4%	7.8%	0.9%
75-79	0.0%	0.3%	0.4%	1.0%	0.4%	0.0%	4.0%	16.2%	22.8%	24.8%	15.8%	3.5%	17.3%	9.2%	1.2%
80-84	0.0%	0.1%	0.6%	0.8%	0.6%	0.1%	1.7%	8.5%	17.7%	19.7%	14.6%	6.4%	14.4%	10.4%	1.5%
85+	0.2%	1.4%	3.2%	4.4%	4.3%	1.0%	1.1%	7.7%	15.5%	22.1%	19.1%	11.3%	15.5%	11.4%	2.3%

Table 2: Share of observations with big errors in table P16: Household type

	Incorporated places						Unincorporated places						Urban areas		
Row Labels	0-499	500-999	1,000-1,999	2,000-4,999	5,000-9,999	10,000+	0-499	500-999	1,000-1,999	2,000-4,999	5,000-9,999	10,000+	2,500-4,999	5,000-9,999	10,000+
Total:	1.4%	0.2%	0.0%	0.0%	0.0%	0.0%	4.0%	0.7%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Family households:	5.8%	2.0%	0.2%	0.1%	0.0%	0.0%	22.9%	18.9%	6.6%	0.9%	0.1%	0.0%	0.4%	0.0%	0.0%
Married couple family	0.4%	0.5%	0.0%	0.0%	0.0%	0.0%	26.1%	37.1%	23.0%	8.5%	1.6%	0.2%	2.6%	0.1%	0.1%
Other family:	4.6%	12.3%	14.8%	7.4%	1.5%	0.2%	11.2%	40.9%	51.9%	42.5%	27.0%	10.1%	19.7%	10.5%	1.8%
Male householder, no spous	0.7%	3.5%	7.6%	16.9%	19.7%	4.9%	1.3%	11.7%	26.6%	40.8%	48.9%	32.6%	30.0%	30.8%	9.1%
Female householder, no spo	0.8%	5.3%	11.0%	7.3%	1.8%	0.2%	6.4%	30.5%	41.9%	42.1%	26.2%	9.5%	22.0%	13.1%	2.1%
Nonfamily households:	3.3%	6.0%	2.7%	1.0%	0.1%	0.0%	22.0%	43.6%	35.0%	17.1%	5.2%	1.3%	5.4%	1.6%	0.2%
Householder living alone	2.2%	4.4%	2.1%	0.4%	0.3%	0.0%	19.7%	41.7%	36.4%	18.9%	6.7%	1.2%	6.0%	2.4%	0.3%
Householder not living alone	0.2%	2.7%	8.4%	20.5%	24.4%	7.3%	2.1%	13.1%	22.0%	35.2%	30.3%	18.1%	30.5%	32.6%	9.4%

Table 3: Share of observations with big errors in table H13: Household size

	Incorporated places						Unincorporated places						Urban areas		
	0-499	500-999	1,000-1,999	2,000-4,999	5,000-9,999	10,000+	0-499	500-999	1,000-1,999	2,000-4,999	5,000-9,999	10,000+	2,500-4,999	5,000-9,999	10,000+
Total:	1.4%	0.2%	0.0%	0.0%	0.0%	0.0%	4.0%	0.7%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
1-person household	2.2%	4.4%	2.1%	0.4%	0.3%	0.0%	19.7%	41.7%	36.4%	18.9%	6.7%	1.2%	6.0%	2.4%	0.3%
2-person household	5.9%	15.7%	6.1%	1.2%	0.0%	0.0%	22.7%	42.2%	27.3%	11.0%	1.7%	0.7%	5.8%	0.7%	0.4%
3-person household	3.0%	15.0%	24.4%	13.2%	3.5%	0.2%	11.2%	34.8%	41.9%	23.2%	4.3%	0.5%	21.9%	7.1%	0.4%
4-person household	2.4%	11.9%	21.6%	18.1%	4.7%	0.4%	8.6%	30.2%	40.9%	28.7%	7.9%	1.5%	27.6%	12.4%	1.1%
5-person household	1.3%	9.4%	18.1%	29.6%	24.5%	4.4%	3.4%	17.0%	27.6%	33.7%	22.4%	6.1%	36.2%	28.6%	6.6%
6-person household	0.3%	3.9%	11.8%	25.7%	41.0%	29.6%	0.5%	3.4%	10.1%	21.6%	31.5%	27.9%	27.9%	41.2%	27.0%
7-or-more-person ho	0.2%	2.9%	8.3%	21.1%	39.8%	44.2%	0.4%	3.1%	7.7%	17.3%	30.1%	39.0%	20.8%	37.0%	38.3%

6 RACE/ETHNICITY OF HOUSEHOLDER

6.1 RESEARCH QUESTIONS:

- Is there significant difference between the race distribution of the householders in SF1 and in the DHC?
- What are the differences in headship and homeownership rates by race?
- What are the differences in race/ethnicity of householders of three generational households?

The Census Bureau presentations for FSCPE and NAC indicated that these questions were investigated at the Census Bureau as well. I think improved accuracy within these tables is very important for housing affordability and access.

Assuming that improvements that were mentioned in the NAC webinar are being implemented, I only show limited results from our own analysis on this topic for the first two questions.

6.2 CONCLUSIONS

The diversity of householders in the demonstration data is often higher than for the same geography in SF1. This is especially true if there was little diversity in SF1.

The calculation of headship rates require data from the persons and data from the households which is problematic and often leads to impossible results.

Diversity among householders in households with three or more generations is rather different between the demonstration data and SF1

6.3 RACE/ETHNICITY DISTRIBUTION OF HOUSEHOLDERS (TABLE H7)

I selected tracts in New York with householders in the demonstration data and in SF1. I than calculated a diversity index based on the 7 Non-Hispanic race groups and the 7 Hispanic race-groups for the demonstration data and for SF1 and looked at the difference between those two.

The tracts were split in three groups based on the SF1 diversity index. The diversity index can be seen as the probability index that two random people are of the same race. A low diversity index means a very homogeneous group and a high index is observed in very diverse geographies.

The mean and mean absolute differences are calculated as well as an indicator which of the two sources indicated more diversity.

Table 4: Difference in tract level diversity between demonstration data and SF1, grouped by SF1 diversity

		Difference between DP and SF1			
Diversity Index value	N	Mean	Mean Absolute	P95 absolute	Prob(DP > SF1)
Low diversity (DI <= 1/3)	2426	2.5%	3.3%	10.0%	70.1%
Medium diversity (1/3 < DI <= 2/3)	1527	0.9%	2.1%	6.3%	62.3%
High diversity (DI > 2/3)	894	0.2%	0.8%	2.6%	53.8%

The mean difference was greater than zero in all three groups indicating that the demonstration data had more diversity among householders than SF1. This is confirmed with the share of observations where the demonstration data had more diversity than SF1.

Another way of looking at this table is by setting thresholds, just like the Census Bureau did for PL95. For 95% of observations the absolute percent error for the largest race of householder category should be less than 5%.

I analyzed the tracts and looked for the minimal population size where the tracts with a larger population had an absolute error of less then 5% in the largest race group. For the tracts in New York State this threshold for the tracts was 2,717 people.

- The absolute percent error for the largest race group within the 3,543 tracts with a population > 2717 had an average of 1.8% and exceeded 5% in 177 tracts (5% of N). The mean percent error was -0.75% indicating a slight negative bias.
- The absolute percent error for the largest race group within the 1,312 tracts with a population <= 2717 had an average of 7.3% and exceeded 5% in 395 tracts (30% of N). The mean percent error was -6.0% indicating a significant negative bias.

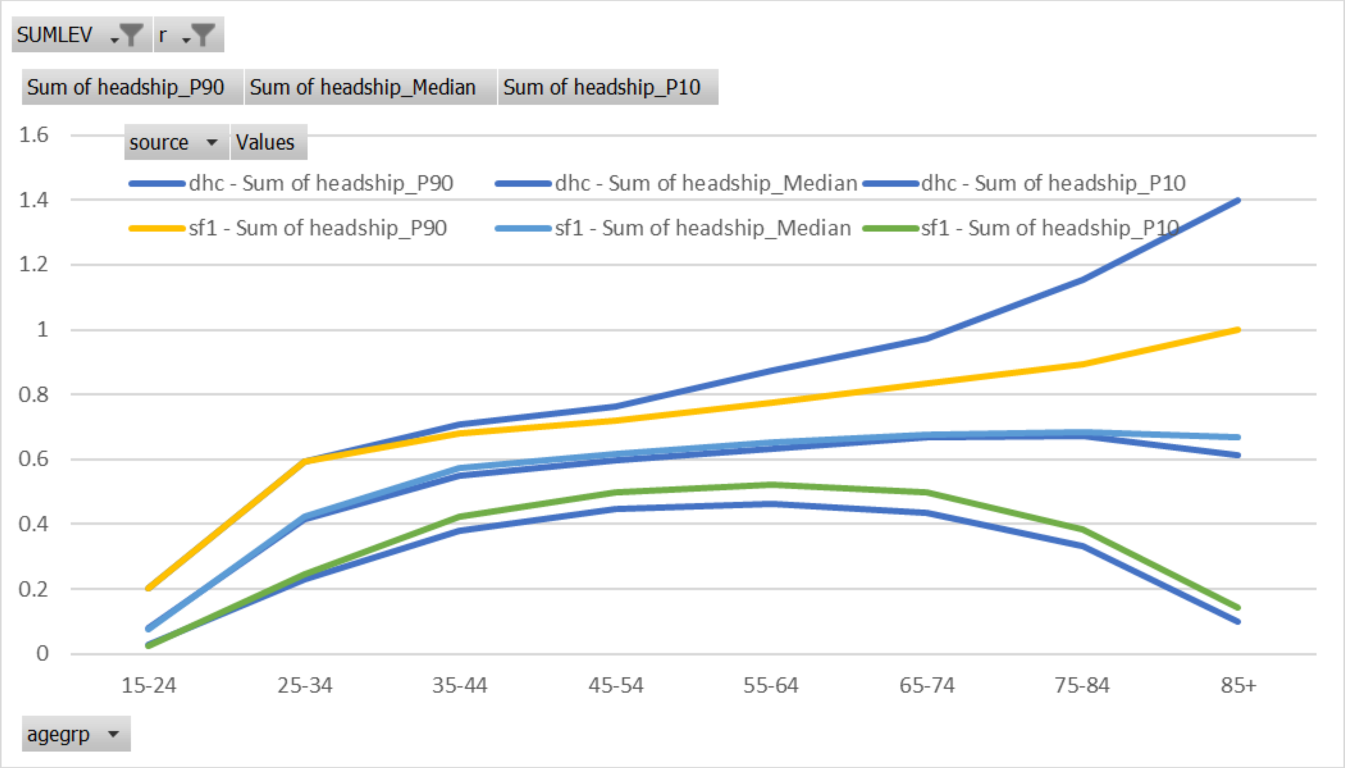
6.4 HEADSHIP AND HOMEOWNERSHIP RATES

I compared the distributions of headship rates (householders in a group divided by household population in that group) and of homeownership rates (homeowners in a group divided by householders in that group) for the several race groups and age groups as presented in table H13 and its iterations.

I compared the 10th percentile, the median and the 90th percentile for each of rates for different levels of geography (with at least 200 householders in the group) between the demonstration data and SF1.

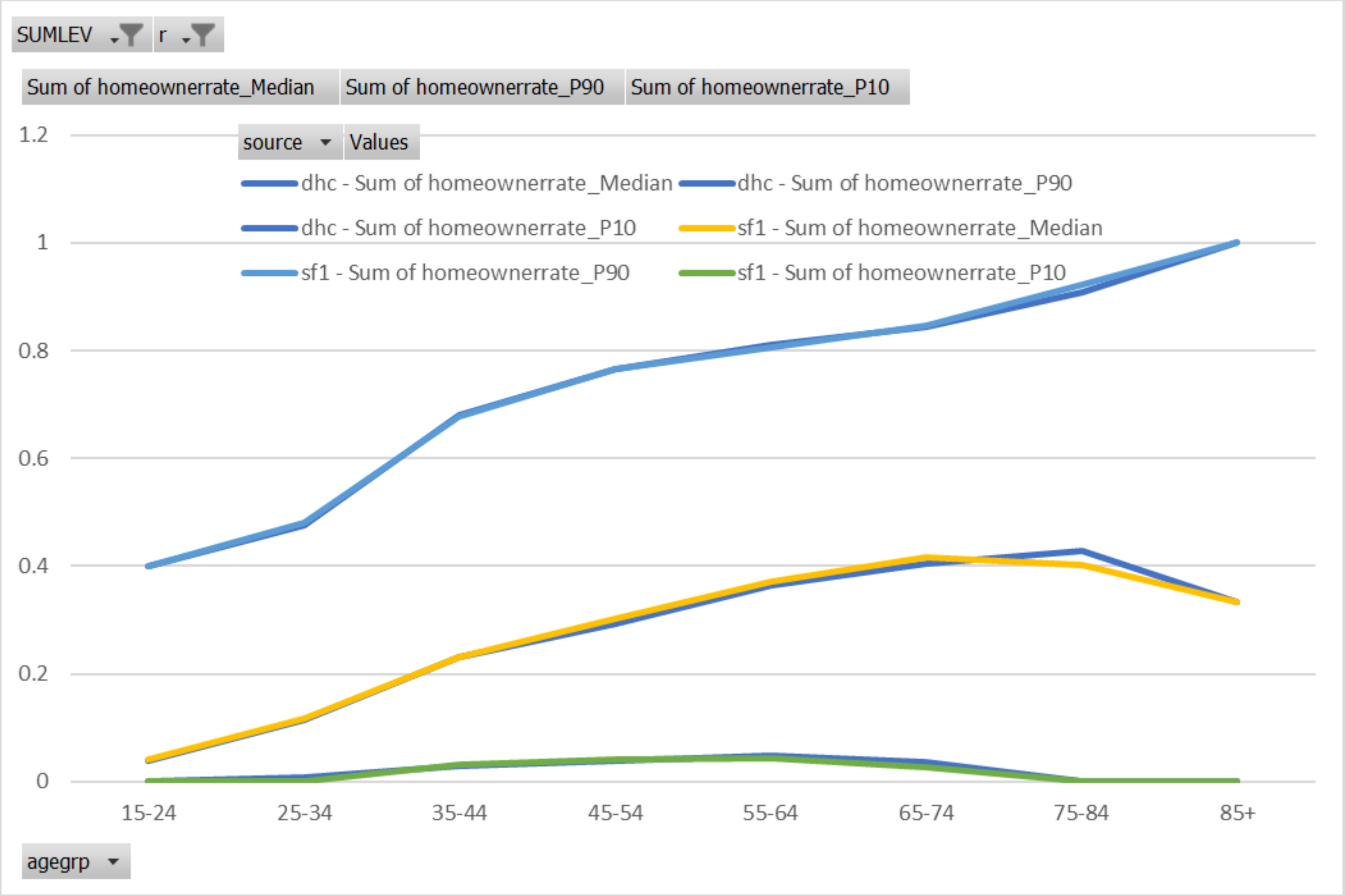
I present here only tract level results for Black householders as they are representative of these analysis.

Figure 22: Distribution of tract level Black Alone headship rates from demonstration data and SF1



The median headship rates are not very different, except for the 85+ population, but the P10 and especially P90 of the headship rate distribution show differences. Theoretically headship rates can not exceed 1, but because population and households are treated independently can be greater than 1 in the demonstration data. For higher ages this happened in more than 10% of tracts and this caused the P90 to exceed 1 in the demonstration data.

Figure 23: Distribution of tract level Black Alone homeownership rates from demonstration data and SF1



The distribution of homeownership rates at the tract level looks very similar between demonstration data and SF1. Please keep in mind that this is not an analysis of the differences at the tract level, only a look at distribution as a result of the aggregate.

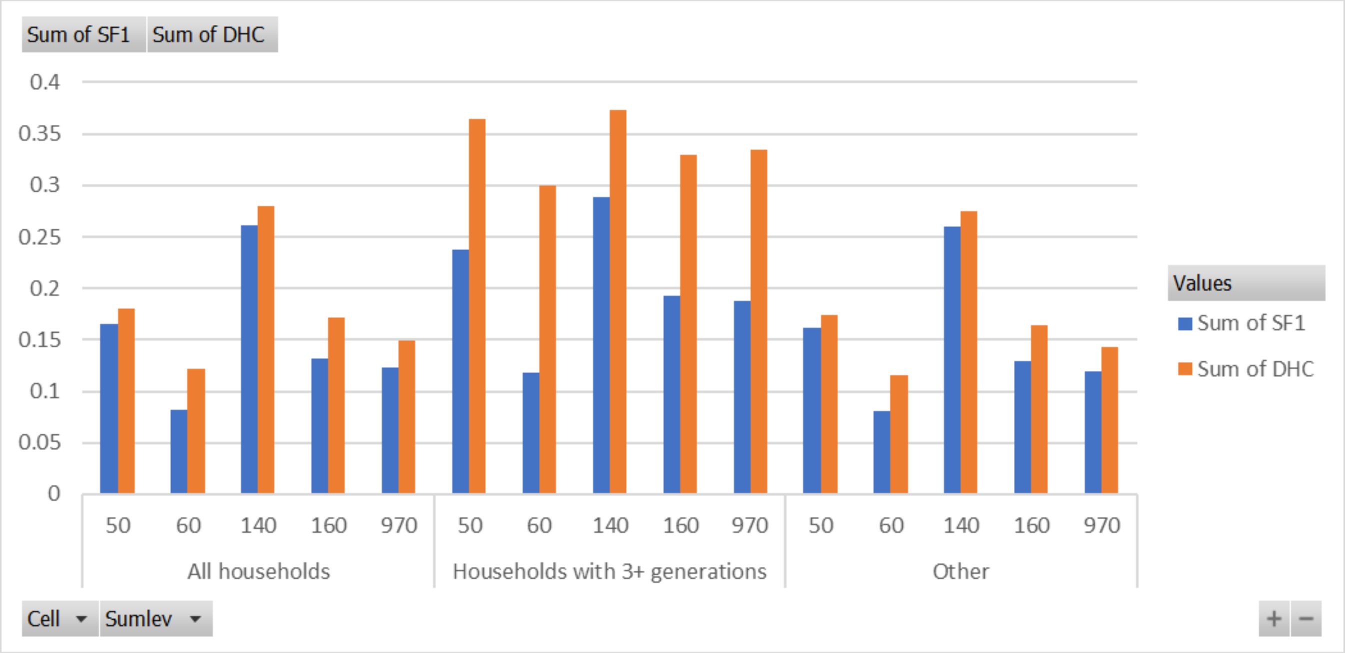
6.5 HOUSEHOLDERS OF THREE GENERATIONAL HOUSEHOLDS

For this analysis we looked at table PCT14: PRESENCE OF MULTIGENERATIONAL HOUSEHOLDS and it’s iterations by the major race groups.

Geographies under consideration are Counties (050), SubCounties (060), tracts (140), places (160) and unified school districts (970) in New York State.

Based on table iterations A through H, I calculated separate diversity indices for all householders and for householders in households with or without the presence of three generations. This diversity index lost a bit of practical interpretation because the groups under consideration are not mutual exclusive, but differences in these diversity indices are still indicative of problems.

Figure 24: Diversity index for householders of households with or without three or more generations



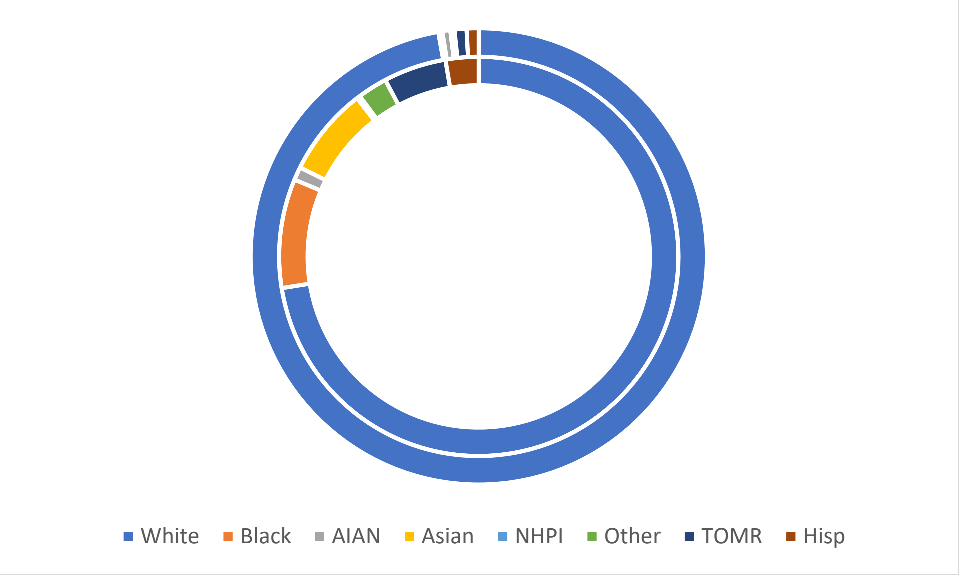
6.5.1 Conclusion

Diversity indices in the demonstration data are higher than in SF1, especially for the rarer category of households with 3 or more generations. This is true for all geographic summary levels considered.

Example: Wyoming County, NY

Outer ring SF1 race of householders of households with 3 or more generations (342 households). Inner ring DHC (271 households)

Figure 25: Race distribution of householders of households with three or more generations in Wyoming County, NY



7 ACCURACY OF THE LARGEST TABLE VALUE

7.1 RESEARCH QUESTIONS:

What can we say about accuracy of the largest value in a table?

In the evaluation of the PL94-171 demonstration data, the Census Bureau set accuracy goals for the largest race/ethnicity groups. This question can be generalized to accuracy of the largest value in a table.

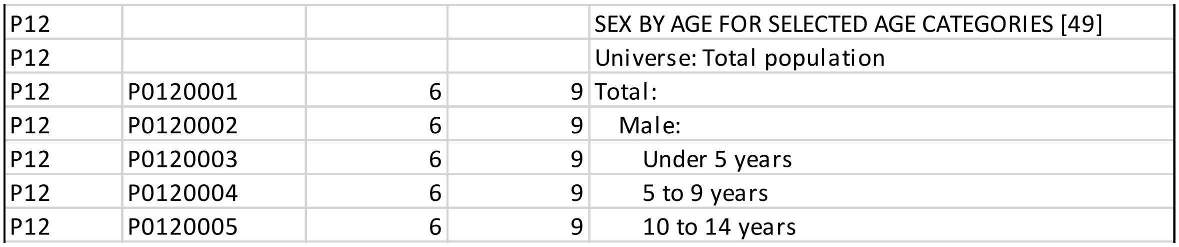
7.2 CONCLUSIONS

The share of the cell with the largest value is on average almost always lower in the demonstration data than in SF1. The larger this maximum value, the smaller the average percentage point difference.

7.3 METHODOLOGY

I used the walkover table to match tables from the demonstration data with SF1 and manually selected as many matches as I could, also based on the number of cells in corresponding tables. No effort was taken to adjust table formats to make even more matches, for example some Group Quarter population by age tables are comparable, but the demonstration data contained fewer cells to limit the numbers of cells with structural zeroes.

Within each table I selected cells that didn't have a further breakdown; no following cells with a larger indent in the walkover table. For example, in the P12 age by sex table the totals by sex are followed by totals by age within this sex that have a bigger indent in the walkover table:



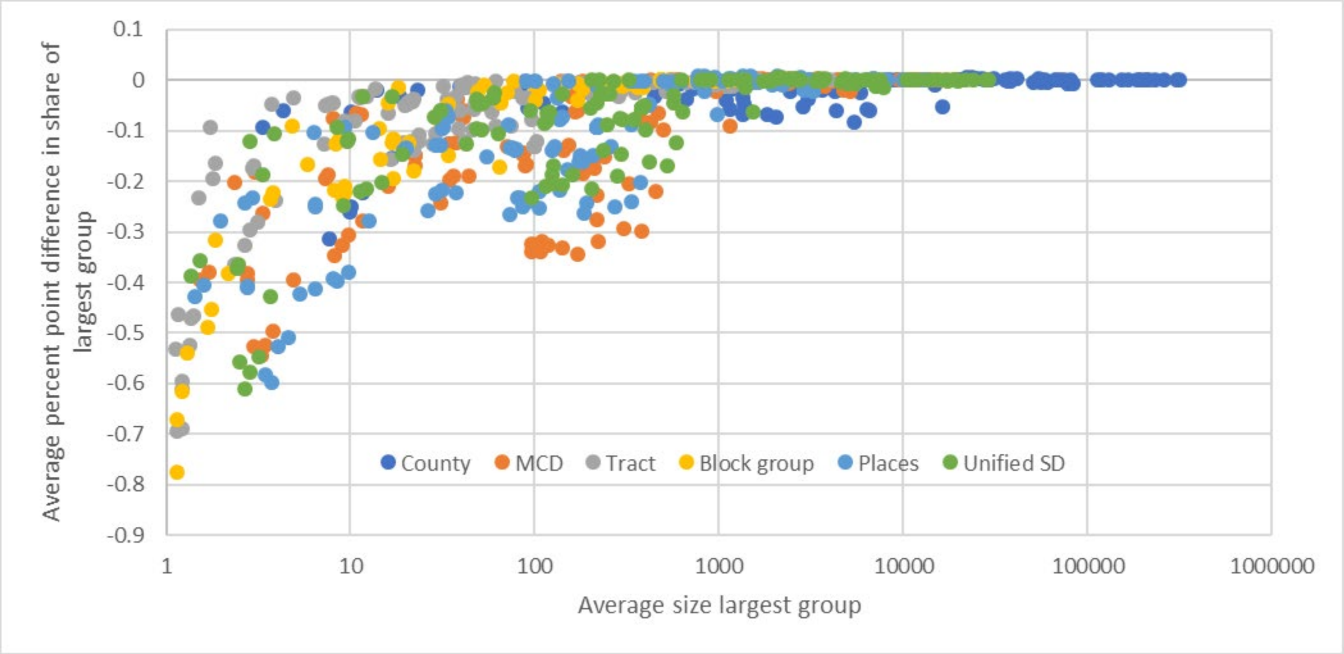
P12				SEX BY AGE FOR SELECTED AGE CATEGORIES [49]
P12				Universe: Total population
P12	P0120001	6	9	Total:
P12	P0120002	6	9	Male:
P12	P0120003	6	9	Under 5 years
P12	P0120004	6	9	5 to 9 years
P12	P0120005	6	9	10 to 14 years

This analyses only used the cells with the biggest indent to find the largest value in the SF1 table. For this cell I calculated a percent error as (DP-SF1)/0.5*(DP+SF1)

For each table and each geography I then calculated an average percent error.

7.4 RESULTS

Figure 26: Average percentage point difference for share of largest table cell



Each dot n the chart represents results from one table and one level of geography. For example the dark blue dot with an X-value of 16229.32 and an Y value of -0.0516 represents County level results for table PCT14B: PRESENCE OF MULTIGENERATIONAL HOUSEHOLDS (BLACK OR AFRICAN AMERICAN ALONE HOUSEHOLDER). The average value of the largest cell in this table is 16,229 and on average the share of this cell is 5.16 percentage point lower in the demonstration data compared to SF1.

Conclusion: The share of the cell with the largest value is on average almost always lower in the demonstration data than in SF1. The larger this maximum value, the smaller the average percentage point difference. It looks like the closer to the spine the smaller the average decrease, but more analyses is needed.

8 DIFFERENCES BY TENURE

8.1 RESEARCH QUESTIONS:

Does the accuracy of the August release of the 2010 Demonstration Data to the original Summary File 1 vary when separated by tenure majority (for example high rental areas compared with high home ownership areas)? Are estimates of households in certain tenure majority areas more vulnerable to inaccuracies than others? Does this depend on the geography of analysis?

8.2 CONCLUSIONS:

- Levels of accuracy varied clearly by tenure, household type, and geography. The most and least accurate tenure type depended on the measure being analyzed and geographic aggregate level.
- Aggregated tracts were consistently more accurate than aggregated block groups for all measures of interest, tenure types, and geographic levels.
- Rental majority areas were highly vulnerable to inaccuracy for households with children and large households (5 or more people)
    - Large households had the highest median absolute percent errors (MdAPEs) of all household types, reaching up to 34.4% in rental majority block groups in Monroe County.
- Errors were significantly smaller for nonfamily households and single person households, but still displayed differential patterns of accuracy.
    - The largest MdAPEs for these household types occurred in owner majority areas across all geographies, except for single person households in New York state where the least accurate tenure type was rental majority.

8.3 METHODOLOGY:

- We used the 2010 Summary File 1 and the August 2022 release of the 2010 Differential Privacy DHC housing unit files at the Census Tract and Block Group levels of geography, excluding Puerto Rico.
- This analysis was limited to census tracts with 200 or more households and block groups with 150 or more households to exclude special use areas.
- Majority housing tenure was determined by calculating percent ownership in a tract or block group: [(IFF002_sf + IFF003_sf)/H8C001_sf] *100
- Geographies were classified into the following categories based on the share of owner-occupied households: rental majority if the share of ownership was <=20%, mixed tenure if the share of owned households was between 21% and 79%, and majority owned if >=80% of occupied households were owned.
- Our final dataset contained 71,842 Census tracts and 214,558 block groups.

8.3.1 Metrics of error:

This section of analysis focuses on one key metric, Median Absolute Percent Error (MdAPE) as a measurement of accuracy to the original Summary 1 File. We utilize the median rather than the mean to be robust to outliers. In our calculations of error, cases where the original SF1 was 0 and the DHC was nonzero resulted in an observation of 0.5.

8.4 Results

8.4.1 Comparing the Accuracy of the 2010 SF1 to DHC: Households with Children (<18)

Chart 1: MdAPE of Census Tracts, by Aggregate Level

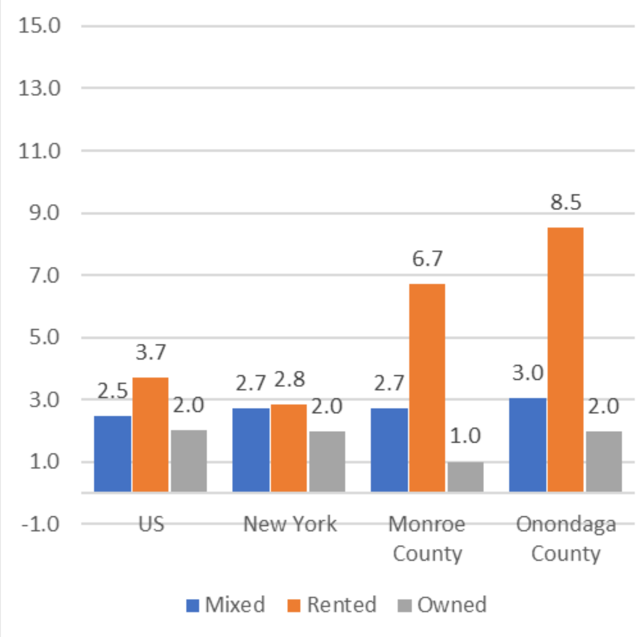
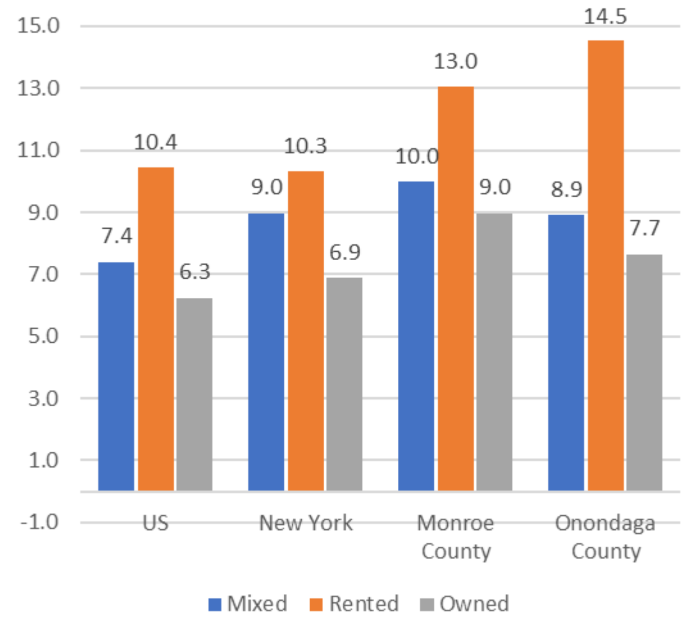


Chart 2: MdAPE of Block Groups, by Aggregate Level



Conclusions for Charts 1 & 2:
-Block group-level counts of households with children under 18 were less accurate to the original SF1 than tract-level counts for all tenure categories and aggregate levels.

- At the tract level, the median absolute percent error (MdAPE) between the original 2010 SF1 and the 2010 DHC for households with children in rental majority areas ranged from 3.7% to 8.5%, while this range at the block group level was from 10.4% to 14.5%.

- Among census tracts, the MdAPE for owner majority and mixed tenure areas were low, ranging from around 1% to 3%, but still rose to between 6% and 10% at the block group level.

- At both the tract and block group level, counts of households with children under 18 in rental majority areas were less accurate to the original SF1 than owner majority and mixed tenure areas.

Estimates of households with children in owner-majority areas were consistently the most accurate at every aggregate level (i.e. national, state, and county).

8.4.2 Comparing the Accuracy of the 2010 SF1 to DHC: Non-family households

Chart 1: MdAPE of Census Tracts, by Aggregate Level

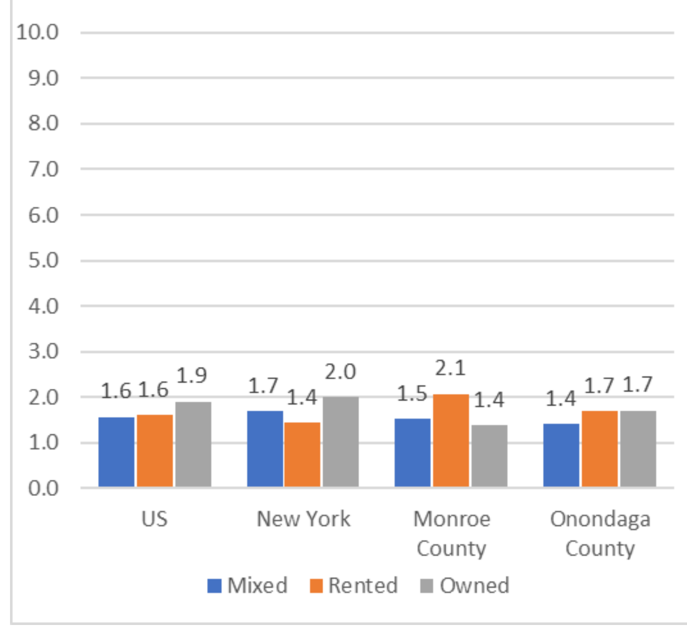
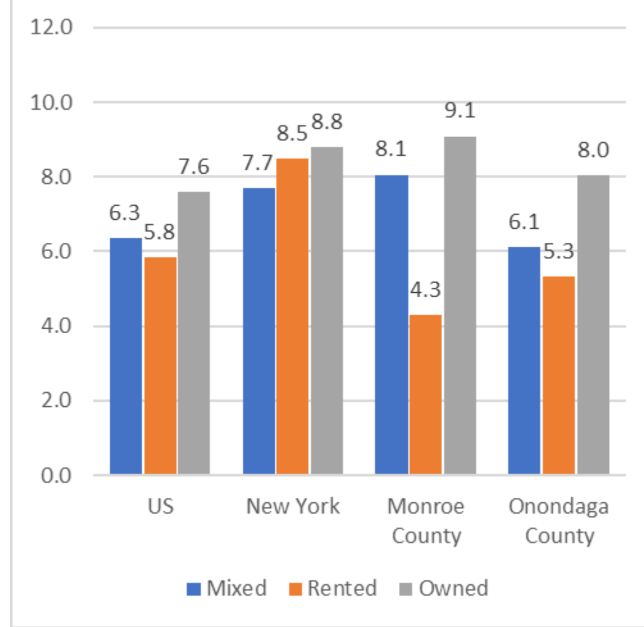


Chart 2: MdAPE of Block Groups, by Aggregate Level



Conclusions:

-Though block groups remained less accurate than tracts, patterns among non-family households by tenure and geographic level differed from those observed for large households and households with children.

- At the tract level, differences between the files remained low. However, owner majority areas had the highest MdAPE for all geographies except for Monroe County, where the largest MdAPE was found among nonfamily households in rental majority areas.
- Owner majority areas at the block group level were the least accurate for all geographies, ranging from 7.6% in the U.S. to 9.1% in Monroe County. For block groups, households in rental majority areas of Monroe County had the lowest MdAPE.

- For nonfamily households, block groups aggregated to the New York State level were less accurate at all tenure areas than nationally aggregated block groups. Conversely, Onondaga County block groups were more accurate than New York State for all tenure types.

8.4.3 Comparing the Accuracy of the 2010 SF1 to DHC: Single Person Households

Chart 1: MdAPE of Census Tracts, by Aggregate Level

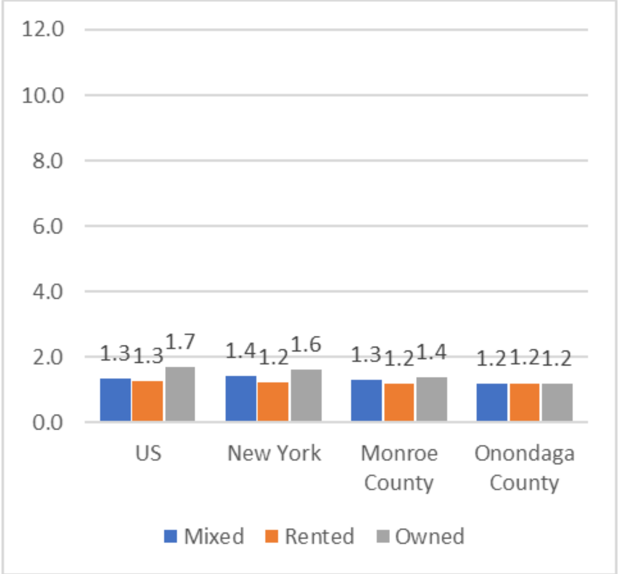
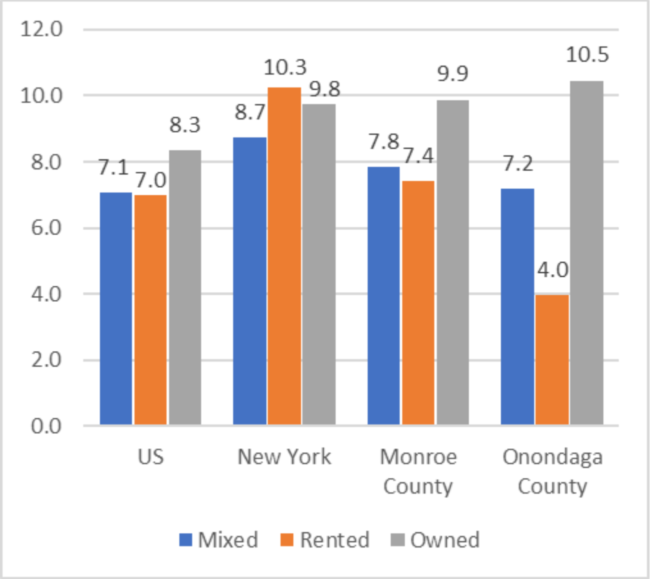


Chart 2: MdAPE of Block Groups, by Aggregate Level



-Patterns of accuracy for single person households mirrored those of nonfamily households.

- Errors among aggregated tracts were consistently low.

- Single person households in owner majority block groups were the least accurate for most geographies, ranging from 8.3% in the U.S. to 10.5% in Onondaga County.

    - State level aggregated block groups were the exception, with rental majority areas producing the highest MdAPE (10.3%). For all other geographies, single person households in rental majority areas were the most accurate.

As with nonfamily households, block group estimates of single person households in New York State were less accurate at all tenure areas than at the national level.

- Both Onondaga and Monroe County had lower MdAPEs than New York State for single person households in mixed and rental majority areas.

- MdAPEs of single person households in owner majority areas were greater at the county level than the state level.

8.4.4 Comparing the Accuracy of the 2010 SF1 to DHC: Large Households (5+ People)

Chart 1: MdAPE of Census Tracts, by Aggregate Level

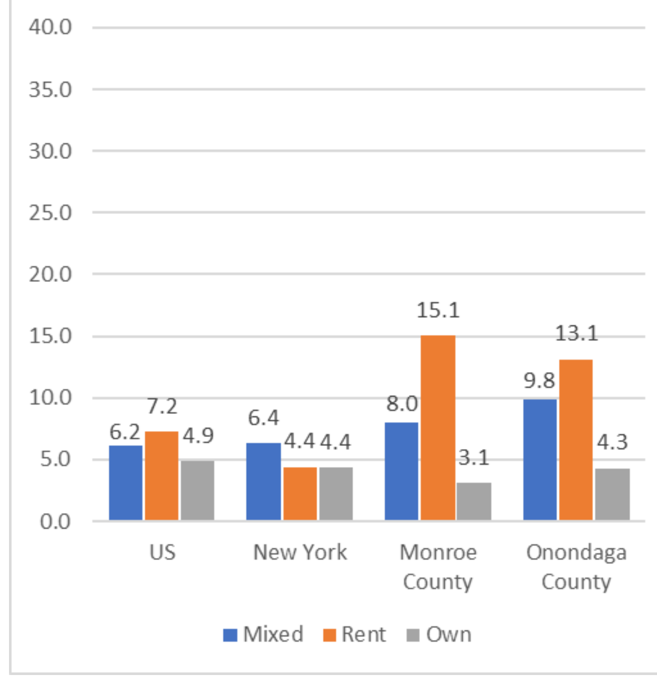
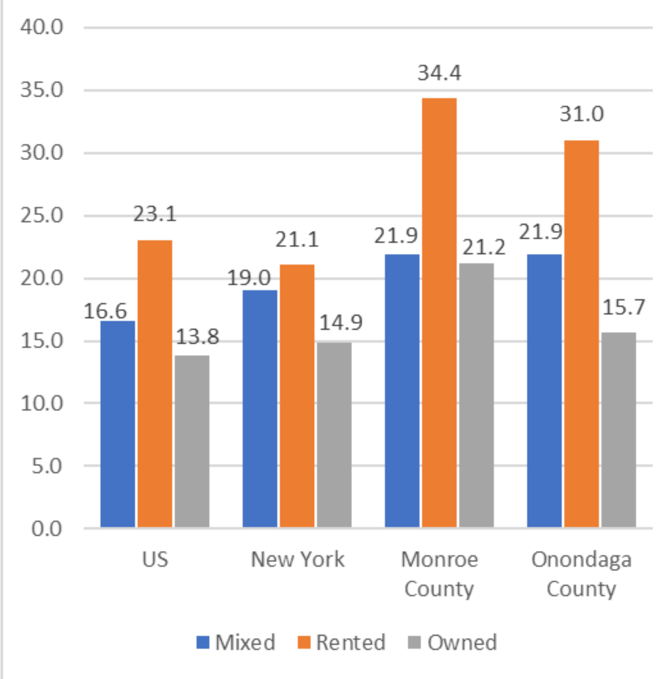


Chart 2: MdAPE of Block Groups, by Aggregate Level



Conclusions:

-Of the four variables examined in this analysis, estimates of large households in the DHC were the least accurate to the original SF1.

-At the tract level, the median absolute percent error (MdAPE) between the original 2010 SF1 and the 2010 DHC for large households in rental majority areas ranged from 4.4% to 15.1%, while this range at the block group level was from 21.1% to 34.4%.

- Among census tracts, the MdAPE for owner majority areas ranged from 3.1% to 4.9%, but still rose to between 13.8% and 21.2% at the block group level.
- Similarly, counts of large households in mixed tenure areas had MdAPEs between 6.2% and 9.8% at the tract level and between 16.6% and 21.9% at the block group level.

- Though large households in rental majority areas had the highest MdAPE across all geographies, this was not the case at the tract level.

- For large households in tracts aggregated to New York State, mixed tenure areas were the least accurate while renter and owner majority areas had the same MdAPE.

9. Erica Maurer (DCP), Senior Demographic Analyst – Population, NYC Department of City Planning

Hello,

Please accept the attached letter on behalf of New York City Department of City Planning as feedback for the 2020 Census Data Products—August 25, 2022 demonstration data for the 2020 Census demographic and housing characteristics file.

Thank you,
Erica Maurer
Senior Demographic Analyst • Population
NYC Department of City Planning

September 23, 2022
RE: 2020 Census Data Products

On behalf of the New York City Department of City Planning (DCP), I am pleased to respond to the August 25, 2022 request for feedback on the *2020 Census Data Products*.

As we did with the March 2022 DHC demonstration data, we have limited our assessment of the August 2022 DHC demonstration data to variables that are really critical to our operations. The most important component of the DHC, from our perspective, is the 5-year age sex breakdown, because this is the key input for our population projections and estimates. We evaluated the 5-year age sex data by census tract and for our geographic unit of analysis, Neighborhood Tabulation Areas, or NTAs, which are rough approximations of New York City neighborhoods built out of census tracts. Our finding is that the latest release showed improvements, even over the March 2022 data which we already deemed fit for use. We saw that in general, there was a closer match in the slopes of the data and diminished absolute percent differences. However, tracts are still showing problematic deviations, beyond 10%, in certain age groups, even if limited to small tract population sizes. Users should be aware of this limitation and urged to use the data with caution and aggregate to larger geographic areas.

We are very appreciative to see that the 5-year age-sex data for the household population was included below the tract level since the last release. These data are essential as an input to the population projections at the neighborhood-level using the cohort component model. In our review of the data, these data were also evaluated at the tract and NTA levels and were deemed fit for use, with the same caveat that we saw some problematic deviations, beyond 10%, in certain age groups.

To reiterate from previous feedback submissions, a great concern is the reduced geographic specificity associated with detailed race and Hispanic Origin data – the current proposal offers detail down to the county-level, whereas the 2010 Census had detail down to a census tract-level, which is crucial to our understanding of the nuance underlying race and Hispanic ethnicity. In New York City, it is not enough to know, for example, that the Asian population has decreased in Manhattan’s Chinatown. We must disentangle subgroup information by race, distinguishing whether it was the Chinese or Vietnamese population that declined in this example, so that we can properly allocate resources for services that our residents require. It is important to consider that even when an overall race group remains unchanged, we may still see significant ethnic transitions among detailed racial subgroups. The Census Bureau has invested years of work towards improving and expanding the race and Hispanic questions, so that we can more precisely portray our increasingly diverse population. For the 2020 Census, the Bureau collected roughly 350 million write-in responses across all racial and Hispanic ethnicity groups, compared to about 50 million collected in 2010. However, the current product plan does not do justice to this collection effort, and respondent burden, because it fails to tap the rich responses that can better portray the diversity of neighborhoods across the nation.[1]

[1] In New York City, these current proposals to reduce critical 2020 Census data detail from the census tract level up to the county-level will be particularly damaging, as our smallest county has nearly a half million people, while our average tract has a population of about 4,000. Consequently, the current proposal will reduce geographic detail for key characteristics by more than 100-fold. Unfortunately, the same can be said for many locales across the country.

In summary, the DHC data we evaluated seem fit for use and we are greatly appreciative that variables deemed crucial to our work were now included at lower levels of geographies. In the same regard, we strongly recommend that the detailed race and Hispanic Origin data be released down to a tract-level.

Sincerely,
Erica Maurer
Senior Demographic Analyst
Population Division

10. Louise Rollin, Alamillo Los Angeles County Department of Public Health, Office of Health Assessment and Epidemiology

Dear DHC Data Products Team –

Please see attached a Case Use Comment Letter regarding the proposed DHC Disclosure Avoidance System (DAS) tuning as provided in the DHC Data Demonstration data released on August 25, 2022. This letter was signed by Dr. Paul Simon, Chief Science Officer of the Los Angeles County Department of Public Health.

We have also attached a document which include two case uses. Please see attached Case Uses DHC Data Demonstration LAC DPH OHAE.docx. The attached excel document, Table 2.xlsx, accompanies and is referenced in the Case Uses…OHAE.docx document.

If you could let me know that you are receipt of this letter, that would be appreciated.

Please let us know if you have any questions about our case uses and how they were developed.

We eagerly await to hear the proposed outcomes regarding the release of this vitally important data.

Sincere Regards,

Louise Rollin-Alamillo, MS
Chief Research Analyst
Los Angeles County Department of Public Health
Office of Health Assessment and Epidemiology

COUNTY OF LOS ANGELES
Public Health

BARBARA FERRER, Ph.D., M.P.H., M.Ed.
Director

MUNTU DAVIS, M.D., M.P.H.
County Health Officer

MEGAN McCLAIRE, M.S.P.H.
Chief Deputy Director

313 North Figueroa Street, Suite 806
Los Angeles, CA 90012
TEL (213) 288-8117 • FAX (213) 975-1273

PAUL SIMON, M.D., M.P.H.
Chief Science Officer
313 North Figueroa Street, 6th Floor-West, Room 610
Los Angeles, California 90012
TEL (213) 288-7280 FAX: (213) 202-2161

www.publichealth.lacounty.gov

Robert L. Santos, MA
U.S. Census
Bureau
4600 Silver Hill
Road Washington,
DC 20233

September 26, 2022

Dear Mr. Santos:

I am writing to urge the US Census Bureau to use a disclosure avoidance system (DAS) that minimizes the error introduced into the census-tract level estimates in the forthcoming *Census 2020 Demographic and Housing Characteristics* file *(DHC).* As Chief Science Officer for the Los Angeles County Department of Public Health, I am concerned that the level of error contained in the current DHC demonstration file may negatively impact our Department's ability to measure and address health inequities if it is retained in the final DHC file, as it will not allow us to access highly accurate population estimates by broad race and ethnicity categories at sub-county geographies for LA County.

Given the vast size and diversity of our county's population as well our Department's deep commitment to identifying and eliminating health inequities within our communities, having access to accurate race and ethnicity data at sub-county geographies is critical for us and our many partners in examining and understanding the myriad of factors that impact the health and well-being of our residents. This is especially true for certain communities with relatively smaller population sizes but who are disproportionately afflicted by adverse health outcomes, including our county's American Indian/Alaska Native (AIAN) and Native Hawaiian/Pacific Islander (NHPI) populations (two priority populations for our Department). To provide a recent example, we have relied on having access to accurate census-tract level race and ethnicity population data throughout the COVID-19 pandemic to understand the disproportionate impact of the pandemic on our county's AIAN and NHPI communities. These data have also been used extensively to inform our COVID-19 vaccination efforts, from estimating coverage rates in various subpopulations by city and community to informing our targeted outreach efforts.

In addition, the DHC has historically served as the foundation for the detailed postcensal population estimates that are procured by the Los Angeles County government for several of its departments, including the Department of Public Health. These postcensal estimates necessarily provide highly detailed demographic information at the split census tract-level (stratified by detailed age groups, race and ethnicity, and sex). Our Department relies on these postcensal estimates as a source of information during intercensal years for the demographic composition of the county overall and various sub-county geographies as well as a source of denominator data to use for our incidence and prevalence rate calculations. Errors in the DHC file will therefore be carried forward for the remainder of the decade when these subsequent postcensal estimates are produced.

We are concerned that if the level of error introduced by the current tuning of the DAS in the DHC demonstration data remains in the final DHC file, the subsequent population data that we frequently use to characterize or county populations at various sub-county geographies or as denominators in our incidence and prevalence rate estimates could be distorted. This would make it difficult to ascertain overall accuracy or whether any observed increases or decreases in trends are real or attributable to data artifact (especially for smaller populations). This could in turn hamper our Department’s surveillance, evaluation, and intervention efforts or result in inappropriate resource allocation.

In summary, I urge you to utilize a DAS that will minimize the amount of error contained in the final *DHC* file as these data are vitally important to public health efforts in our county.

Sincerely,

Paul

Paul Simon, MD, MPH
Chief Science Officer
Los Angeles County Department of Public Health

PS:AK

Table 2. Age adjusted Mortality Rates for all causes of death, by race/ethnicity and sex, Service Provider Areas (SPAs) Los Angeles County, 20
Published Census 2020 population estimates vs. estimates adjusted by Disclosure Avoidance System (DAS)*****

Race Ethnicity	SPA 1			SPA 2			SPA 3			SPA 4			SPA 5	
	SF1	DAS-adj	% change	SF1	DAS-adj	% change	SF1	DAS-adj	% change	SF1	DAS-adj	% change	SF1	DAS-adj
White Total	932.6	930.9	0%	633.3	634.6	0%	709.8	707.8	0%	697.6	697.1	0%	515.3	517.4
Black Total	1012.2	1036.1	2%	791.3	747.0	-6%	843.1	841.5	0%	764.2	757.3	-1%	699.9	696.9
Asian Total	405.1	385.9	-5%	437.1	427.6	-2%	405.3	408	1%	459.7	460.3	0%	372.3	371.9
Latino Total	542.8	536.9	-1%	468.8	470.2	0%	563.3	561.6	0%	535.0	545.0	2%	443.4	440.5
AIAN Total	*	*	*	654.5	719.3	***10%***	946.5	921.5	-3%	1214.6	1295.5	7%	*	*
NHOPI Total	*	*	*	*	*	*	*	*	*	*	*	*	*	*
White Males	1092.7	1079.2	-1%	748.9	751.5	0%	847.3	845.6	0%	793.8	793.4	0%	571.9	574.9
Black Males	1156.5	1227.4	6%	863.7	811.7	-6%	966.4	951.6	-2%	874.6	853.5	-2%	842.4	837.4
Asian Males	558.4	539.7	-3%	525.7	506.6	-4%	479.0	484.2	1%	648.8	651.9	0%	480.6	479.2
Latino Males	638.0	627.2	-2%	559.3	558.3	0%	689.0	685.5	-1%	659.6	673.8	2%	542.0	546.4
White Females	794.0	798.3	1%	537.7	538.3	0%	602.8	600.6	0%	598.1	597.7	0%	465.8	467.2
Black Females	906.5	912.6	1%	727.2	687.9	-5%	751.9	756.6	1%	649.0	648.3	0%	589.0	590.7
Asian Females	317.1	300.8	-5%	376.2	371.6	-1%	347.2	348.6	0%	332.9	332.6	0%	300.2	300.3
Latino Females	463.8	461.7	0%	400.6	403.1	1%	471.9	471.0	0%	433.5	441.5	2%	366.7	360.7

** # of deaths are less than 20 and omitted because unreliable.*

*** Summary File 1 (SF1) population estimates are from the PCT12I-M series*

**** Demographic and Housing Characteristics (DHC) flie adjusted with Disclosure Avoidance System (DAS) and released as Data Demonstration on August 25, 2*

10.

	SPA 6			SPA 7			SPA 8		
% change	SF1	DAS-adj	% change	SF1	DAS-adj	% change	SF1	DAS-adj	% change
0%	1280.7	1157	***-10%***	795.9	793.2	0%	727.3	726.1	0%
0%	1029.3	1043.1	1%	779.2	683.9	***-12%***	886.3	885.6	0%
0%	543.4	567.5	4%	416.5	413.6	-1%	468.5	472.3	1%
-1%	480.0	468.8	-2%	564.0	568.7	1%	505.1	503.5	0%
*	*	*	*	902.6	816.1	***-10%***	772.4	772.9	0%
*	*	*	*	*	*	*	1006.1	1111.5	***10%***
1%	1637.6	1453.0	*-11%*	975.5	965.5	-1%	842.4	840.8	0%
-1%	1306.3	1333.0	2%	1089.9	933.9	***-14%***	1092.8	1093.8	0%
0%	752.5	827.2	*10%*	500.6	500.6	0%	561.7	562.8	0%
1%	560.1	543.2	-3%	691.1	697.8	1%	616.3	620.6	1%
0%	982.0	894.6	-9%	654.6	654.7	0%	624.2	623.2	0%
0%	840.2	849.1	1%	571.5	505.6	***-12%***	742.3	741.1	0%
0%	379.2	380.6	0%	351.6	347.7	-1%	400.9	406.0	1%
-2%	409.4	401.1	-2%	465.9	469.6	1%	414.7	410.6	-1%

022. PCT12I-M Series

Case Uses for the DHC Data Demonstration Released August 25, 2022

Office of Health Assessment, Los Angeles County Department of Public Health

This document includes two case uses and a section on data analysis findings based on our review of the Summary File 1 and the DHC Data Demonstration Data.

Case Use 1: Covid-19 Vaccination Coverage. The DHC data used for this analysis were the P12I, P12H, P12J, and P12L tables.

Background: As part of the work to evaluate Covid-19 vaccination efforts, our Department produces reports of vaccination coverage for different race/ethnicity groups and age groups for Countywide Statistical Areas (CSA) which are comprised of split census tracts. Percentages of persons in groups stratified by race/ethnicity and age groups that have received vaccination are reported as well as a corresponding indication of the group's vaccination range as either "below", "equivalent", or "above" the County's vaccination level.

To produce an assessment of how the 2010 SF1 data in tables P12H (Hispanic), P12I (White, Non-Hispanic), P12J (Black Non-Hispanic), and P12L (Asian Non-Hispanic) compared with these tables adjusted with Disclosure Avoidance System (as released on August 25, 2022), we decided to derive how vaccination coverage ranges would change by using the DAS estimates as population denominators. This was done with the following steps: 1) we selected a subset of 150 CSAs that were comprised of whole tracts or the CSA had the overwhelming of the census tract's population (if the tract was shared with another CSA); 2) We re-calculated *hypothetical* ranges of persons who received a Covid-19 vaccine for each of the 150 selected countywide statistical areas (CSAs) in Los Angeles County (LAC) for persons over the age of 65 years and for 4 race/ethnicity groups (White NH, Asian NH, Black NH, and Hispanic). This was done by first multiplying the percentage range of persons vaccinated with the population estimate using the published SF1 estimates to get an estimated range of persons vaccinated. Secondly, we divided this derived range of persons that were vaccinated with the population denominators for that group that came from the DAS adjusted SF1 estimates. This led to the calculation of new percentage range of persons that were vaccination. From this new estimated range, we could then identify if the County's comparison assessment was the same or if it had changed with the use of the new population denominators.

As an example, the population estimates of Latinos 65 years and older in the Crenshaw District CSA increased from 196 (SF1) to 280 (SF1 adjusted by DHC). The range of Covid-19 vaccination coverage for that CSA in 2022 is 93% to 100% for this group, which is considered "Equivalent" to its County counterpart range. When we calculate 93% of 196 and 100% of 196, we got an estimate that approximately 182 – 196 persons that were vaccinated. When we took that range and divided it with the population estimate from the SF1 adjusted by DAS. This produced a new percentage range of persons vaccinated – 65% (182 divided by 280) to 70% (196 divided by 280). This new percentage range was no longer "Equivalent" to the County's range (which for this CSA was >95% vaccination coverage) but has fallen to "Below' the County's range. In this hypothetical case, the rates changed because of the population change of the SF1 to the adjusted SF1 (increase of 43%).

Findings: Of the 150 CSAs, 16 would have hypothetically changed their county assessment category for at least one of the race ethnicity groups based only on the change in the population denominators. These were for race and age stratified groups with 100 or more persons. These results are hypothetical as we are using 2010 population estimates (those of the SF1 and adjusted by DAS) as the population denominators for Covid-19 vaccination coverage rates which are currently calculated with 2019 population estimates.

We also calculated the mean absolute percent errors (MAPEs) for persons 65 years and older by the 4 race groups and across all 150 CSAs (see table 1 below). As the table shows the percent errors for the selected CSA was greatest among Blacks over the age of 65 years, followed by Asians, followed by Whites, and followed lastly by Latinos.

Table 1. Ranking of Mean Absolute Percent Errors* for persons 65 years and older by specified Race/Ethnicity, 150 selected cities and communities, LA County, 2010.**

Rank	Race/ethnicity and Age group	Mean Absolute Percent Error***
1	Blacks, 65+	19.8%
2	Asians, 65+	16.2%
3	Latinos, 65+	12.5%
4	Whites, 65+	12.0%

** Errors between 2010 Summary File 1 Estimates and 2010 SF1 estimates and 2010 DHC estimates.*

*** CSAs comprised of only whole 2010 census tracts or tract assigned to CSA based on very large majority of tract population was in assigned CSA, if tract was split with other CSAs.*

**** This is a measure of the "average" absolute difference for a particular statistic. For example, for the population of Asians over the age of 65 years at the CSA Level, calculate [Abs(DAS-adjusted - SF1)/SF1] for each of the 150 CSAs, then take the mean.*

<u>Case Use 2: Age-Adjusted Mortality Rates by Race/Ethnicity for Service Provider Areas.</u> The DHC data used for this analysis were the PCT12I, PCT12H, PCT12J, PCT12L, PCT12K, and PCT12 M tables.

In this case use, age-adjusted mortality rates (AAMRs) for all causes of death were calculated using both the published estimates as denominators (SF1) and the SF1 estimates adjusted by DAS. These were derived by race/ethnicity for each Service Provider Areas (SPAs). There are eight SPAs that cover Los Angeles County and they are comprised of census tracts.

When comparing the derived rates, we found that rates varied by more than 10% for the following groups (see Attachment Table 2.xlsx): Blacks in SPA 7 (-12%), American Indian and Alaska Native (AIAN) in SPA 1 (10%) and in SPA 7 (-10%), Native Hawaiian and Other Pacific Islander (NHOPI) in SPA 8 (10%), and lastly for Whites in SPA 6 (10%).

When we stratified the rates further by sex, the ranking by AAMR in males changed from first highest to second highest for Black men in SPA 7, as that rate decreased by 14% when using the DAS adjusted population estimates in place of the published SF1 rates.

Data Analysis of the DHC: Mean Absolute Percent Errors (MAPES) calculated for race/ethnicity, sex, and age stratification among LAC's census tracts.

We calculated the mean absolute percent errors (MAPEs) for population estimates stratified by age, sex, and race/ethnicity using both the published 2010 Summary File 1 estimates and the DHC Data Demonstration data set. This was done using the Tables of P12 A-I for Los Angeles County's 2,346 census tracts.

Table 3 shows the mean absolute percent difference for each race and ethnicity group by age. The two highest mean absolute percent errors for each category are highlighted. The darker orange is the highest mean absolute percent error and the lighter orange is the second highest mean absolute percent difference.

The group with the largest mean absolute percent difference for Total Population is the NHOPI alone group (56.1%), followed by the AIAN alone group (14.6%). In regard to sex, the group with the highest mean absolute percent difference for the male group is the NHOPI alone group (male = 67.0%), followed by the AIAN alone group (male = 32.5%). The group with the highest mean absolute percent difference for the female group is the NHOPI alone group (male = 67.2%), followed by the AIAN alone group (34.0%). Amongst almost all the race/ethnicity groups, the MAPEs calculated by age and sex are either first or second highest for the American Indian Alaska Native (AIAN) or Black population.

Table 3: Mean Absolute Percent Error*, for groups stratified by sex, age, and race/ethnicity and across Census Tracts, Los Angeles County, 2010

	White	Black	AIAN	Asian	NHOPI	Hispanic	White, NH
Total	0.6%	4.0%	14.6%	4.4%	56.1%	1.4%	1.6%
Male	1.6%	11.1%	32.5%	11.6%	67.0%	1.5%	5.4%
Under 5 years	9.5%	55.5%	64.5%	43.0%	33.0%	14.2%	38.5%
5 to 9 years	8.3%	50.5%	63.5%	38.4%	31.3%	12.5%	34.5%
10 to 14 years	8.0%	49.7%	66.4%	38.2%	33.8%	11.4%	34.8%
15 to 17 years	14.2%	62.6%	64.2%	50.0%	29.6%	20.1%	47.6%
18 and 19 years	16.4%	60.4%	59.3%	53.9%	27.0%	20.7%	46.4%
20 years	37.3%	62.8%	42.1%	62.8%	20.1%	36.9%	58.6%
21 years	37.0%	66.4%	43.5%	63.7%	17.1%	41.2%	57.8%
22 to 24 years	16.5%	61.0%	67.2%	46.8%	34.1%	21.7%	40.4%
25 to 29 years	8.6%	54.7%	79.5%	36.5%	45.3%	13.7%	27.4%
30 to 34 years	9.5%	54.8%	75.4%	37.0%	42.7%	14.0%	32.9%
35 to 39 years	8.8%	51.8%	74.8%	36.4%	41.2%	13.4%	31.7%
40 to 44 years	8.6%	48.7%	78.2%	36.2%	42.0%	12.8%	29.1%
45 to 49 years	8.4%	48.6%	75.3%	36.4%	38.8%	13.3%	27.0%
50 to 54 years	8.9%	47.9%	78.1%	34.5%	34.9%	16.7%	25.9%
55 to 59 years	9.9%	53.5%	74.0%	38.0%	31.4%	19.5%	25.6%
60 and 61 years	21.5%	67.0%	39.6%	50.1%	14.2%	38.6%	37.2%
62 to 64 years	19.4%	64.1%	49.1%	48.7%	16.8%	35.9%	34.0%
65 and over	9.9%	65.4%	74.3%	38.3%	38.9%	20.9%	26.1%
Female	1.9%	12.7%	34.0%	9.5%	67.2%	1.6%	5.9%
Under 5 years	9.1%	56.7%	67.9%	41.4%	31.3%	13.0%	34.9%
5 to 9 years	8.0%	50.3%	62.6%	39.4%	28.1%	12.4%	33.3%
10 to 14 years	8.5%	48.0%	64.8%	37.5%	30.6%	11.8%	33.9%
15 to 17 years	15.2%	61.6%	69.0%	53.2%	33.6%	19.6%	48.3%
18 and 19 years	17.7%	60.9%	55.0%	54.9%	23.5%	22.8%	50.6%
20 years	38.1%	64.5%	38.2%	65.4%	15.5%	41.0%	59.0%
21 years	38.9%	60.5%	39.3%	64.7%	16.3%	41.4%	57.8%
22 to 24 years	16.7%	62.3%	65.8%	48.3%	30.6%	22.0%	40.0%
25 to 29 years	9.4%	52.2%	75.8%	34.4%	43.5%	13.8%	32.9%
30 to 34 years	8.9%	52.2%	77.5%	34.2%	43.5%	12.6%	33.6%
35 to 39 years	8.8%	50.9%	75.9%	33.6%	38.3%	11.5%	34.5%
40 to 44 years	16.4%	51.4%	79.4%	32.6%	41.2%	12.1%	31.3%
45 to 49 years	8.9%	45.5%	78.2%	33.3%	38.7%	12.2%	30.1%
50 to 54 years	8.8%	48.2%	77.7%	32.4%	33.8%	14.6%	27.7%
55 to 59 years	31.9%	54.7%	74.2%	35.4%	28.5%	16.5%	28.2%
60 and 61 years	22.1%	63.7%	45.7%	48.8%	17.1%	35.7%	39.3%
62 to 64 years	17.9%	63.1%	52.2%	45.9%	20.5%	31.2%	36.6%
65 and over	9.1%	63.4%	81.0%	34.2%	40.1%	16.8%	25.8%

** This is a measure of the "average" absolute difference for a particular statistic. For example, for the population of Asians over the age of 65 years at the CSA Level, calculate [Abs(DAS-adjusted - SF1)/SF1] for each of the 150 CSAs, then take the mean.*
***AIAN is American Indian Alaska Native and NHOPI is Native Hawaiian Other Pacific Islander.*

11. Kim, Yang-Seon, Research and Statistics Officer, State of Hawaii

Hi,

This is Yang-Seon Kim, the FSCPE and FSCPP representative for the state of Hawaii. I summarized the findings from my review of the demonstration data for Hawaii in the Word file attached here. Detailed comparisons I did to write the summary are included in the three Excel files attached here.

If possible, can you send me an email acknowledging your receipt of this email?   I hope that my review helps in improving the quality of the 2020 data.

Thanks,

Yang-Seon Kim, Ph.D.
Research and Statistics Officer
Research and Economic Analysis Division
Department of Business, Economic Development & Tourism
State of Hawaii

ED. NOTE: SPREADSHEET ATTACHMENTS NOT INCLUDED DUE TO SIZE LIMITATIONS. CONTACT 2020DAS@CENSUS.GOV TO REQUEST AN EMAILED COPY.

Review of the 2010 demonstration data for Hawaii

By Yang-Seon Kim (FSCPE representative for Hawaii, yang-seon.kim@hawaii.gov)

Sep 26, 2022

This document summarizes what we found in the review of the 2010 demonstration data for Hawaii released by the Census Bureau on Aug 25, 2022. The review was done based on the demonstration data prepared by the IPUMS NHGIS.

Detailed comparison of DP (Privacy-Protected demonstration data) with SF (original 2010 data from the Summary File 1) for each variable of our interest for four geographic areas (state, counties, census tracts, and block groups) are included in the three attached Excel files.

Total population and households

We learned that the data were modified even for total variables such as total population, total household population, and total number of households in the area. The size of the percent differences was mostly determined by the size of the area: small percent differences for the areas with a big population and big percent differences for the areas with a small population.

At the census tract or block group level, there were many areas where the total numbers were significantly modified. Table 1 shows some examples of the areas with a big percent difference between SF and DP. We have a big concern about this kind of modification as these total numbers are what governments and other various data users rely on to make decisions on the area.

Other than the size of the area, we didn't find any factor that caused a systematic difference in the size of the percent differences between SF and DP.

Table 1. Some examples of the areas with a big difference between SF and DP value for **total** population and total households

GEOCODE	Level	Variable	Name	SF	DP	Abs. (DP-SF)	Abs. (% dif)
150090307062	BG	H7V001	Total population	18	37	19	51.4%
150070403005	BG	H7V001	Total population	274	394	120	30.5%
150010216041	BG	H7V001	Total population	382	505	123	24.4%
150070403005	BG	H8A001	Total household population	250	368	118	32.1%
150010216041	BG	H8A001	Total household population	382	505	123	24.4%
150030098012	BG	H8A001	Total household population	725	895	170	19.0%
150070405001	BG	H8C001	Total number of households	93	119	26	21.8%
150030098012	BG	H8C001	Total number of households	292	260	32	12.3%
150030093003	BG	H8C001	Total number of households	272	247	25	10.1%

CT: census tract, BG: Block group
SF: Original 2010 data from Summary File 1
DF: Privacy-Protected 2010 Census Demonstration Data

Detailed population and household variables

The patterns we found in the size of percent differences between SF and DP for some detailed population and household variables were similar to those of total variables; mostly determined by the size of the universe of the variable. Table 2 shows some examples of the areas with a big percent difference between SF and DP for some population and household variables that we frequently use for a study or policy decision making. There were a large number of areas where the modification was very significant.

Table 2. Some examples of the areas with a big difference between SF and DP value for some detailed population and household variables

GEOCODE	Le	Variable	Name	SF	DP	Abs	Abs
	vel					(DP-SF)	(% dif)
Population variable							
150030098012	BG	H8A002	Under 18 in households	143	248	105	42.3%
150030073031	BG	H8A003	18 + in households	2	23	21	91.3%
15003009704	CT	H81001	Total Pop in group quarters	194	126	68	54.0%
150030053001	BG	H81001	Total Pop in group quarters	40	21	19	90.5%
150030089242	BG	H7X002	White alone	143	114	29	25.4%
150010212022	BG	H7X005	Asian alone	252	200	52	26.0%
150030013003	BG	H7X006	NHOPI alone	52	24	28	117%
150010216041	BG	H7X008	Two or more races	26	110	84	76.4%
150070403005	BG	H7X008	Two or more races	50	130	80	61.5%
150030064023	BG	H7X008	Two or more races	308	210	98	46.7%
150010204002	BG	H76002	Total male population	163	244	81	33.2%
150010216041	BG	H76002	Total male population	192	267	75	28.1%
150070403005	BG	H76002	Total male population	151	205	54	26.3%
Household variables							
150030037006	BG	H8C002	Family households	163	108	55	50.9%
150010203001	BG	H8C002	Family households	207	153	54	35.3%
150010203001	BG	H8C003	Family HH-husband & wife	158	103	55	53.4%
15003007701	CT	H8C004	Family HH -other family	321	244	77	31.6%
15003009603	CT	H8C005	Family HH-Male hholder, no wife	227	149	78	52.3%
150030080031	BG	H8C007	Nonfamily HH	97	187	90	48.1%
150030080031	BG	H8C008	Living alone	64	131	67	51.1%
150070405001	BG	IFF001	Total occupied Housing units	93	119	26	21.8%
150030098012	BG	IFF001	Total occupied Housing units	292	260	32	12.3%
150070405001	BG	IFG006	Vacant: for seasonal, recreational….	28	4	24	600%
150030018012	BG	IFG006	Vacant: for seasonal, recreational….	57	22	35	159%
150090314053	BG	IFH002	Householder, White alone	113	42	71	169%
15003010201	CT	IFH005	Householder, Asian alone	128	175	47	26.9%
150030094003	BG	IFH005	Householder, Asian alone	210	124	86	69.4%
150010212022	BG	IFH005	Householder, Asian alone	83	197	114	57.9%
150030086063	BG	IFH006	Householder, NH alone	102	58	44	75.9%
150030106022	BG	IFH008	Householder, two more races	114	66	48	72.7%
150030086063	BG	IFH008	Householder, two more races	142	86	56	65.1%

CT, BG, SF, DF – see the footnote of Table 1

Median age

We reviewed the median age of nine population groups. For the groups we reviewed, the size of the differences depended mostly on the size of the universe (population that the median age was calculated for), and the difference was unacceptably big when the universe was very small such as less than 50 people in the group. The maximum difference between SF and DP median age we observed for each population group was in the range of 30-60 years.

Although the differences tended to decrease with the size of the universe, more than 10 years of difference in two median ages was easily found in the groups with population bigger than 100. The following table shows some examples of the areas with relatively big difference in SF median age and DP median age. For the block group presented in the top row of the table, the DP median age was 13 years lower than the median age in SF even though the area had almost 400 people in 2010.

We don't know whether we want to use median ages and other age data for small geographic areas if these scales of noises are added to the 2020 Census data.

Table 3. Some examples of the areas with a big difference between SF and DP median age

GEOCODE	Level	Variable	Population that the median age was calculated for and the population in 2010		Median age		
					SF	DP	Abs. (DP-SF)
150010216041	BG	H77001	Total population, both sexes	382	56	43	13
150030008004	BG	H77001	Total population, both sexes	468	59	67	8
150030099021	BG	H77001	Total population, both sexes	648	43	36	7
150030038003	BG	H77001	Total population, both sexes	1,006	63	56	6
150070404004	BG	H77001	Total population, both sexes	1,304	44	38	6
150030013003	BG	H77002	Total population, male	468	35	47	11
150030112013	BG	H77002	Total population, male	411	50	39	11
15007040104	CT	H9P001	Two or more races, both sexes	230	23	32	9
150030060002	BG	H9P001	Two or more races, both sexes	141	46	29	16
150010204002	BG	H9P001	Two or more races, both sexes	107	35	23	12
150030046001	BG	H9P001	Two or more races, both sexes	179	30	18	12
15007040104	CT	H9P002	Two or more races, male	120	19	37	18
15003003407	CT	H9P002	Two or more races, male	51	37	18	19
150030001073	BG	H9P002	Two or more races, male	42	17	38	21
150030034063	BG	H9P002	Two or more races, male	135	29	17	12
150030064011	BG	H9P002	Two or more races, male	123	29	18	11
150070407006	BG	H9M001	Asian alone, both sexes	93	55	44	11
150030008004	BG	H9M001	Asian alone, both sexes	253	72	82	10
150030086141	BG	H9M001	Asian alone, both sexes	150	40	31	9
15009030403	CT	H9M002	Asian alone, male	160	50	39	11
150030034061	BG	H9M002	Asian alone, male	189	41	55	14
150030111041	BG	H9M002	Asian alone, male	101	48	61	13
150010217042	BG	H9M002	Asian alone, male	121	43	31	12
150030087012	BG	H9M002	Asian alone, male	123	43	55	12

CT, BG, SF, DF – see the footnote of Table 1

12. Abraham D Flaxman Associate Professor, Institute for Health Metrics and Evaluation | University of Washington

linked_census_disclosure_2022_09_26.pdf

Dear DAS Team,

Attached please find an update to our investigation into disclosure risk around transgender identity from a reconstruction-abetted linkage attack. We updated our methods slightly from our previous version and applied them to your new demonstration data product, which yielded largely similar results.

--Abie

THE RISK OF LINKED CENSUS DATA TO TRANSGENDER CHILDREN: A SIMULATION STUDY

ABRAHAM D. FLAXMAN AND OS KEYES

Institute for Health Metrics and Evaluation, University of Washington
e-mail address: abie@uw.edu

Abstract. Every ten years the United States Census Bureau collects data on all people living in the US, including information on age, sex, race, ethnicity, and household relationship. They are required by law to protect this data from disclosure where data provided by any individual can be identified, and, in 2020, they used a novel approach to meet this requirement, the differentially private TopDown Algorithm.

We conducted a simulation study to investigate the risk of disclosing a change in how an individual's sex was recorded in successive censuses. In a simulated population based on a reconstruction of the 2010 decennial census of Texas, we compared the number of transgender individuals under 18 identified by linking simulated census data from 2010 and 2020 under alternative approaches to disclosure avoidance, including swapping in 2020 (as used in the 2010) and TopDown in 2020 (as planned for the actual release).

Our simulation assumed that in Texas 0.2% of the 3,095,857 children who were under the age of 8 in the 2010 census were transgender and would have a different sex reported in the 2020 census, and 23% would reside at the same address, which implied that 1,424 trans youth were at risk of having their gender identity disclosed by a reconstruction-abetted linkage attack. We found that without any disclosure avoidance in 2010 or 2020, a reconstruction-abetted linkage attack identified 657 transgender children. With 5% swapping in 2010 and 2020, it identified 605 individuals, an 8% decrease. With swapping in 2010 and TopDown in 2020 as configured in the August 25, 2022 demonstration data release, it identified 196 individuals, a 68% decrease from swapping.

In light of recent laws prohibiting parents from obtaining medical care for their trans children, our results demonstrate the importance of disclosure avoidance for census data, and suggest that the TopDown approach planned by Census Bureau is a substantial improvement compared to the previous approach, but still risks disclosing sensitive information.

Introduction

As part of the 2020 decennial census, the US Census Bureau has developed a new approach to disclosure avoidance, based on differential privacy, called the TopDown Algorithm (TDA) (Abowd et al. [2019]). The details of their approach have been refined iteratively since they first debuted as part of the 2018 end-to-end test (Garfinkel et al. [2019]). The release of the Demographics and Housing Characteristics (DHC) data in 2023 will be the next application of TDA for a data product from the 2020 decennial census. As of September, 2022, we

Key words and phrases: census, disclosure avoidance, linkage attack.

have the products of the first application of TDA (the Public Law 94-171 redistricting data, released in August, 2021) as well as a demonstration DHC product from a test run in March 2022 (Bureau [2022b]) as well as a test run in August 2022 (Bureau [2022a]) to help us understand plans and trade-offs for some of the TDA options previously enumerated (Petti and Flaxman [2019]).

In support of their work to develop and validate TDA, the Census Bureau has previously released a series of Privacy-Protected Microdata Files (PPMFs) by applying iterations of TDA to the 2010 Census Edited File. The DHC products from March and August of 2022 diverge from this pattern and provides summary tables without releasing a corresponding PPMF. This invites the question of whether the release of a PPMF or reconstruction of microdata from DHC tables might compromise privacy. In this work, we investigated empirically how well TDA protects against disclosure of sensitive information on an individual's gender identity in DHC data.

Past investigations of demonstration products have focused primarily on the impact of TDA on accuracy of key census-derived statistics, and we agree that there are broad, political implications behind statistical accuracy; the framing of census data informs everything from the shape and number of legislative districts to funding and resourcing for minority groups (see, for example, Thompson [2012]). But this is also true of privacy—accurate representation is not an unalloyed good. For many groups, particularly those who are vulnerable to and have experienced active discrimination by state entities, higher accuracy can also mean higher identifiability and higher *scrutiny*. An example of this is undocumented immigrants' relation to questions about citizenship—questions that can be used to identify, surveil, and punish people who are undocumented, and consequently lead to reduced engagement with and trust of the census (see Barreto [2019]). More recent in the public eye (although just as longstanding, as highlighted by Canaday [2009]) are questions of gender (Singer [2015]), on which this investigation is focused.

The last few years have seen heightened scrutiny of transgender (henceforth "trans") people, with a particular focus on (and moral panic around) trans children (see Slothouber [2020]). This has included actions by state actors to simultaneously legislate against access to care and equal treatment, and use existing mechanisms of government to punish the children and parents who have become identifiable. Most prominently, the governor of Texas, in Abbott [2022], has directed the state Department of Family and Protective Services to investigate the parents of any trans child who receives gender-affirming medical care. In order to do so, he advocates drawing on existing systems for child and parent surveillance, including abuse reporting requirements, to identify targets.

As all of this suggests, there are many reasons for us to be cautious around data availability and the pursuit of accuracy as an untrammelled good. While it is beneficial from a statistical perspective, an absence of privacy simultaneously risks both producing real, material harms for the individuals identified, and undermining trust in the census itself and so (paradoxically) reducing the very accuracy that is aimed for. To demonstrate the importance of factoring identifiability into account—and the necessity of an emphasis on disclosure avoidance in census policy—we used simulation to investigate a risk to privacy, by focusing on the risk of disclosing a child's transgender status, through discordant reporting of binary gender in successive censuses.

Methods

We used computer simulation to compare the number of trans children who might be identified in a synthetic population under alternative scenarios of disclosure avoidance. Our approach began with a synthetic population of size and structure similar to the state of Texas, derived from a reconstruction of the US population on April 1, 2010. Since our focus is on linking youth between the 2010 and 2020 Decennial Censuses, we included simulants from this population who were aged zero to seven and therefore would be under 18 on April 1, 2020. We augmented this reconstruction by assigning the simulant's gender based on responses to the Sexual Orientation and Gender Identity (SOGI) module of the Behavioral Risk Factors Surveillance System (BRFSS) collected in 2019 (National Center for Chronic Disease Prevention and Health Promotion, Division of Population Health [2019]).

We initialized each simulant with attributes for age, gender, race, ethnicity, and household, where age was an integer value representing the age in years, gender was a five-valued variable (with values of transgender boy; transgender girl; transgender, gender nonconforming; cisgender boy; and cisgender girl), race was a 63-valued variable encoding the possible combinations of the six Census racial categories, ethnicity was a two-valued variable for Hispanic/non-Hispanic, and household was an identifier that encoded census geography (state, county, tract, block) as well as housing unit id.

From this initial population, we simulated the progression of time and the data captured in the 2010 and 2020 Decennial Censuses as follows: we recorded the age at initialization precisely for each simulant's reported age in the 2010 census, and then that age plus 10 for each simulant's reported age in the 2020 census. We used a simple model of the other key demographic factors of births, deaths, in-migration, and out-migration to simulate how this population might change over the next decade: since our interest was in linking between censuses, we focused on migration, and posited that every household might move, making it harder to link. To realize this household mobility, we selected households to stay unmoved from 2010 to 2020 independently, with probability value $p_{\text{stay}} = 23\%$ derived from the American Communities Survey (we obtained this value by calculating the sample-weighted proportion of with-children households that had been in residence for at least 10 years in the 2020 5-year Public Use Microdata Sample). We updated the 2020 address of each non-staying household by selecting a new household for them to move to uniformly at random from all synthetic households in Texas that were occupied on Census Day 2010, which offered a simple way to include realistic heterogeneity in population density.

Finally, we simulated the reported value of sex on the 2010 and 2020 Decennial Census. Our model of reported sex started from the assumption—uncertain though it is—that, in the 2010 Census, nearly all of the transgender youth aged zero to seven had their sex reported based on their gender-assigned-at-birth. We then assumed that, for some of the simulants with transgender identities, this would lead to differing responses in the 2020 Census. Based on this premise, we simulated responses on the 2010 and 2020 census according to the following cases: for cisgender boy simulants, we recorded their sex as male in 2010 and 2020, and similarly for cisgender girl simulants we recorded female. For transgender boy simulants, we recorded their sex as female in 2010 and recorded their sex with a value chosen uniformly at random from the set $\{\text{male}, \text{female}\}$ in 2020. Similarly for transgender girl simulants, we recorded their sex as male in 2010 and with a value of female in 2020 with probability 50%. For transgender, gender nonconforming simulants we recorded their sex as the same value in 2010 and 2020, with the value chosen uniformly at random from the set $\{\text{male}, \text{female}\}$.

We recorded race and ethnicity identically in 2010 and 2020, matching the value of the simulant's race and ethnicity attributes.

We compared four alternative scenarios of disclosure avoidance: (1) extreme disclosure where names were published, allowing even households that moved to be linked between censuses; (2) tables with no disclosure avoidance, where names were not published, but there was no effort to swap or otherwise perturb the data in published tables; (3) disclosure avoidance by swapping, where 5% of households were exchanged with another household to protect privacy; and (4) differentially private disclosure avoidance, where the new TDA approach was used to protect against disclosure in published tables. We now describe our method of quantifying how many transgender simulants would have their gender identity revealed in each of these scenarios.

Extreme disclosure (Scenario 1): In this scenario, we assumed that linking on name, age, race, and ethnicity would be able to identify nearly all simulants with discordantly reported values for sex in the 2010 and 2020 censuses. We therefore counted all simulants with differing values reported for sex in 2010 and 2020 to estimate the number of trans youth who would have their gender identity revealed if census microdata including names were released. We hypothesized that this would total in the thousands or perhaps even tens of thousands.

No disclosure avoidance (Scenario 2): In this scenario, we assumed that the only simulants at risk of being identified as trans youth by a reconstructed-abetted linkage attack were those aged seven and younger in 2010 who had a unique combination of age, race and ethnicity in their census block. Furthermore, we assumed that individuals who moved between the 2010 and 2020 censuses would not have their transgender status revealed and even individuals who were exposed by a unique combination of attributes in 2010 and did not move by 2020 *might* not have their transgender status revealed, if in-migration to their census block resulted in them no longer having a unique combination of attributes in 2020. The simulants who had a unique combination of age, race, and ethnicity in their (unmoved) census block in 2010 and 2020 could be linked by deterministic record linkage on these attributes. We therefore identified all simulants who did not move and had a unique combination of attributes in 2010 and also in 2020, and counted the simulants in this group with differing values reported for sex in 2010 and 2020. This constituted our estimate of the number of trans youth who would have their gender identity revealed by a reconstruction-abetted linkage attack if the tables used for reconstruction were published with no disclosure avoidance measures. We hypothesized that this would total in the hundreds.

Swapping for disclosure avoidance (Scenario 3): We approached this scenario similarly to Scenario 2, but instead of using each simulant's geography directly in the reconstruction-abetted linkage attack, we first chose a random subset of simulants to have their reported location swapped to somewhere other than their true location. We achieved this with a simple model analogous to the model of migration described above, where we selected some households to report in a location that is not their actual location independently, with probability $p_{\text{swap}} = 5\%$ (we chose this value as a modeling assumption broadly aligned with the publicly available information about the Census Bureau's approach to disclosure avoidance in the 2010 Decennial Census). For each of the selected households, we chose a reported location to swap in by selecting a household uniformly at random from all synthetic households in Texas on Census Day 2010.

We then identified all simulants who did not appear to have moved, according to their (possibly swapped) reported location in the 2010 and 2020 censuses, who had a unique combination of age, race, ethnicity, and geography attributes recorded in both censuses, and counted the simulants in this group with differing values reported for sex in 2010 and 2020. This constituted our estimate of the number of trans youth who would have their gender identity revealed by a reconstruction-abetted linkage attack if the tables used for reconstruction were protected by swapping. We hypothesized that this total would be five to 10% lower than the total from the no-disclosure-avoidance scenario, and therefore also reveal sensitive information about hundreds of trans youth.

TDA for disclosure avoidance (Scenario 4): We were not able to approach this scenario in a way completely analogous to Scenarios 1-3. Instead of running TDA ourselves on our synthetic data after simulating forward ten years, we used the Census Bureau's DHC demonstration product to generate our estimate of the risk of a reconstruction-abetted linkage attack in this scenario, which requires additional explanation compared to the previous three scenarios. Since this scenario uses the DHC demonstration product, it has new results since our first version of this report.

We began with a reconstruction exercise, to come up with a reconstructed microdata file (ReMF) consisting of a row for each reconstructed individual and columns for the attributes of age, sex, race, ethnicity, and geography for individuals age zero to 17 that was consistent with the tables from the demonstration DHC product that used the TDA for disclosure avoidance. We similarly generated an ReMF from the corresponding SF1 tables published as part of the 2010 Decennial Census (which used swapping to protect against disclosure). Instead of initializing our synthetic population in 2010 and simulating the progression of time, we initialized our synthetic population in 2020, based on the individuals aged 10 to 17 in the SF1 ReMF. We then simulated the *regression* of time, going backwards from 2020 to the 2010 Census Day, when each simulant would be 10 years younger. We applied our migration model to keep the location in 2010 identical to that in 2020 for only a random fraction simulants, governed again by the parameter p_{stay}.

As in the other scenarios, we endowed each simulant with a gender attribute, which we calibrated to match to measurements from the 2019 BRFSS SOGI module. However, in this scenario, we first set the reported sex in 2020 to match that in the SF1 ReMF, and then set the gender attribute and reported sex in 2010 conditional on the reported sex in 2020. This allowed us to use the demonstration DHC as our proxy for the privacy afforded by TDA in 2020 in our assessment of the number of trans youth who would have their gender identity revealed by a reconstruction-abetted linkage attack using data protected by swapping in 2010 and TDA in 2020.

To complete this approach, we identified all simulants who had a unique combination of age, race, ethnicity in their census block in 2010, and identified which of these simulants matched a unique individual aged 10 years older in the DHC ReMF. For each of these simulants, we then compared the reported sex in the 2010 census with the reported sex in the 2020 census. We counted how many of these links were for simulants who were trans youth. We hypothesized that this would be at least an order of magnitude smaller than the total from the swapping-for-disclosure-avoidance scenario.

Results

Our synthetic population included 25,145,561 individual simulants, matching exactly the 2010 population count for Texas. We focused on the simulants aged zero to seven on April 1, 2010, of which we had 3,095,857. Among these simulants, 0.53% were trans, with 0.18%

trans boys, 0.23% trans girls, and 0.12% gender nonconforming (closely matching the the BRFSS values), which led to 0.2% of the simulants having a different value reported for sex in the 2010 and 2020 censuses. Over the ten years of simulation, the majority of households moved at least once and only 23% of simulants resided in the same census block in 2010 and 2020 (closely matching the ACS values). Taking these together implied that in our simulation there were $(3,095,857 \text{ kids}) \times (0.2\%) \times (23\%) = 1,424$ trans youth who were at risk of having their gender identity disclosed by a reconstruction-abetted linkage attack.

We found that in our scenario with extreme disclosure, where individual-level data with linkable names was published (Scenario 1), linking between 2010 and 2020 census data to identify individuals with discordantly reported values for sex would identify over 6,000 trans kids, accounting for 38% of all trans kids in our simulated version of Texas (the remaining 62% were not identified because their reported sex was concordant in both censuses).

In our scenario where tables like those in SF1 or DHC were published precisely as enumerated, without any disclosure avoidance measures applied (Scenario 2), we found that migration and non-uniqueness substantially reduced the number of trans kids who's gender identity was revealed. However, there were still 1,414,929 individuals who were uniquely identified by the age, race, ethnicity, and location in 2010 and 1,766,968 uniquely identified in 2020. In our simulation, a reconstruction-abetted linkage attack in this scenario identified 657 trans kids.

In our next scenario (Scenario 3), we added swapping-based disclosure avoidance to the tables in Scenario 2, and we found that with respect to a reconstructed-abetted linkage attack, swapping acted similarly to a small boost in migration to prevent identifying trans kids. At the 5% swapping level we used in Scenario 3, we found that a reconstruction-abetted linkage attack identified 605 trans kids, an 8% reduction from the number identified in Scenario 2.

Our final scenario is the closest we considered to the approach proposed by Census Bureau in the most recently released demonstration product. In this scenario, we considered protecting the tables released from the 2010 census with swapping and the tables from the 2020 census with TDA (Scenario 4). We found that this afforded substantially more protection than the other scenarios we considered. Because of the alternative route we took to constructing this scenario, we used a different initial population, starting with 3,009,117 simulants ages 10 to 17 on April 1, 2020. We found that TDA was successful in preventing the bulk of the identifications from Scenario 3; in our simulation, a reconstruction-abetted linkage attack identified 196 trans kids when TDA was used for disclosure avoidance on the 2020 tables, a 68% reduction in the number identified when swapping was used in Scenario 3. This is slightly higher than the number identified in our previous iteration of this analysis, where we used the 2022-03-16 version of the demonstration data and identified 170 trans kids.

1. Discussion

Our simulation results demonstrate the magnitude of the threat that a linkage attack designed to identify trans kids might pose. Were Census Bureau to publish microdata on the 2010 and 2020 census (Scenario 1), it would likely identify the transgender status of over 6,000 trans kids in Texas. In the approach underlying the most recent demonstration data, on the other hand, a reconstruction-abetted linkage attack would likely identify the transgender status of only 196 trans kids in Texas. This is yet another demonstration of the need for disclosure avoidance in the Decennial Census.

The bulk of previously published investigations into the quality of TDA demonstration products have compared with published results from the 2010 Census, and often reported differences. But in such comparisons there is an important limitation, because they compare the (published) results of swapping to the (demonstration) results of TDA applied to the unswapped data. Thus the conclusion of such a comparison is typically limited to proving that the noise introduced by TDA is different than the noise from swapping. This investigation turns this limitation into a strength, since a reconstruction-abetted linkage attack between 2010 and 2020 Decennial Censuses *will* be linking data that has been swapped with data that has been protected by TopDown. Modeling in a simulation framework like the approach developed here could potentially also be used in future investigations to more directly compare the noise introduced by swapping to the noise introduced by TDA.

Limitations: There are at least three simplifying assumptions in this simulation model that constitute limitations which might be the focus of future work. First, the migration model is quite simplistic, and it is likely that further investigation could more accurately incorporate determinants of migration; the probability that a households has stayed unmoved between decennial censuses is likely to vary by household income, for example, which is an attribute that we did not include in our simulation, but could potentially add. Second, our simple model of how sex was reported in 2020 census for trans kids could also be more complex, although it is less clear what sources of data could inform adding this complexity. Third, in this work we assumed that race and ethnicity were unchanged between 2010 and 2020 censuses, but it is likely that evolving conceptions of race and ethnicity have led to some recording of differing values for some individuals, and this would result in some reduction in the number of links in a linkage attack. We conjecture that none of these simplifying assumptions have substantially changed the number of trans kids identified in our scenarios, however.

As mentioned in the methods section, our approach to Scenario 4 is more complicated than we would have liked. However, the approach we used in this work has a strength alluded to above, because it uses SF1 data that has been perturbed by the swapping approach actually employed by Census Bureau in the 2010 Decennial Census, the details of which are not publicly available.

We would also like to emphasize three limitations specific to our model of trans children. First, our assumption of a uniform probability of markers changing between census years is overly-simplistic; we would expect that, in practice, the likelihood of changes is variable depending on both the respondent family's context and the individual perspective of the child and their parent(s). Second, the limited range of sex options on the census means that many trans children whose identities fall outside a simplistic binary do not alter their census markers. Third, we would expect differences in the amount of geographic mobility and consistency in household structure for trans families writ large. While one response to increasing scrutiny, at least for those with means, is to purposefully move their household, it seems logical that in some cases the risk of increased scrutiny will lead others to purposefully *not* move. The latter of these considerations could suggest this is in fact the *minimal* count of trans people identifiable through the current census approach to data disclosure, and that without changes to the data disclosure approach, well-intended efforts to increase the ability of Census Bureau instruments to record and represent trans people (see White House [2022]) could increase the risk of identifiability and harm.

Although the focus of this piece is on trans *children*—specifically, those under 18 in both the 2010 and 2020 census, with different sex records in each—it is worth emphasising that they are not the only people at risk. With the addition of more census tranches (say, 2000, or, going forward, 2030), the range of people at risk of disclosing their transgender status would expand to include trans adults, many of whom, if they have children, are also being targeted for additional scrutiny by state bodies.

Due to data limitations, we had to use computer simulation to conduct this investigation, but it would be possible for Census Bureau to replicate and expand on analyses such as this one internally, where they can use private data such as the Census Edited File, which is not available to outside researchers. The Census Bureau could reproduce this analysis using its internal unprotected data to understand how its implementation differs from this model. We encourage them to share with us how much this risk differs in the true implementation from the risk as modeled in this simulation.

We have made a replication archive of this work available online: https://github.com/aflaxman/linked_census_disclosure

Acknowledgment

The authors wish to acknowledge fruitful discussions with danah boyd and Simson Garfinkel.

References

G. Abbott. Letter to Masters, 2022. URL https://www.documentcloud.org/documents/21272649-abbott-letter-to-masters.

J. Abowd, D. Kifer, B. Moran, R. Ashmead, P. Leclerc, W. Sexton, S. Garfinkel, and A. Machanavajjhala. Census TopDown: Differentially private data, incremental schemas, and consistency with public knowledge. Technical report, U.S. Census Bureau, 2019.

M. A. Barreto. Expert Testimony: *NYAG New York v. U.S. Immigration and Customs Enforcement*, 2019. URL http://mattbarreto.com/papers/Declaration_of_Matthew_A._Barreto_-_NY.pdf.

U. C. Bureau. 2010 demonstration data for the Demographic and Housing Characteristics file (DHC) (v. 2022-08-25). Technical report, 2022a. URL https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/02-Demographic_and_Housing_Characteristics/2022-08-25_Summary_File/.

U. C. Bureau. 2010 demonstration data for the Demographic and Housing Characteristics file (DHC) (v. 2022-03-16). Technical report, 2022b. URL https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/02-Demographic_and_Housing_Characteristics/2022-03-16_Summary_File/.

M. Canaday. The straight state. In *The Straight State*. Princeton University Press, 2009.

S. Garfinkel et al. 2018 end-to-end test disclosure avoidance system design specification. Technical report, U.S. Census Bureau, 2019.

National Center for Chronic Disease Prevention and Health Promotion, Division of Population Health. 2019 brfss survey data and documentation. Technical report, 2019. URL https://www.cdc.gov/brfss/annual_data/annual_2019.html.

S. Petti and A. Flaxman. Differential privacy in the 2020 US census: What will it do? Quantifying the accuracy/privacy tradeoff. *Gates Open Research*, 3, 2019.

T. B. Singer. The profusion of things: the "transgender matrix" and demographic imaginaries in US public health. *Transgender Studies Quarterly*, 2(1):58–76, 2015.

V. Slothouber. (de) trans visibility: moral panic in mainstream media reports on de/retransition. *European Journal of English Studies*, 24(1):89–99, 2020.

D. Thompson. Making (mixed-) race: census politics and the emergence of multiracial multiculturalism in the United States, Great Britain and Canada. *Ethnic and Racial Studies*, 35(8):1409–1426, 2012.

White House. FACT SHEET: Biden-Harris administration advances equality and visibility for transgender Americans, 2022.